Universidad Carlos iii de Madrid

Escuela Politécnica Superior

Ingeniería técnica de Telecomunicaciones

Sistemas de Telecomunicación

PROYECTO FINAL DE CARRERA

Extracción Robusta de Parámetros utilizando Modelos Auditivos Mejorados y Transformaciones Morfológicas para Reconocimiento Automático de Habla

Autor: Josue Igual Blanco
Tutores: Francisco j. Valverde Albacete
Carmen Pelaez Moreno

Julio de 2008

Proyecto Fin de Carrera

Extracción robusta de parámetros utilizando modelos auditivos mejorados y transformaciones morfológicas para reconocimiento automático de habla

Autor
Josué Igual Blanco

Tutores
Francisco J. Valverde Albacete
Carmen Pelaez Moreno

La defensa del presente Proyecto Fin de Carrera se realizó el día 23 de Julio de 2008, siendo evaluada por el siguiente tribunal:

Presidente: Ana I. García Moral

Secretario: Rubén Solera Ureña

Vocal: Luis Antonio Puente

Las interminables horas de prácticas, las semanas de exámenes o las clases soporíferas.
Lo acabé todo, y lo hice con vosotros.
Gracias chicos por todos los ratos geniales entre clases, las broncas en los trabajos en grupo, las cosas de la suerte antes de los examenes y las cervezas de aprobar.
Aunque ahora que vamos acabando seguimos viendonos quería deciros que todos estos años fueron geniales gracias a vosotros,
a todos y todas los de la uni, al one-two-three-four, y muy especialmente a Paula.
Y como no...a toda mi familia, miles de gracias.
Un gran beso a todos.

CORTINILLA DE ESTRELLAS

Introducción

La parametrización de la señal de voz es uno de los pasos fundamentales en el Reconocimiento Automático de Habla.

En la actualidad coexisten diferentes técnicas para realizar esta función. Sin embargo, los porcentajes de error siguen siendo elevados.

Del deseo, y la necesidad, de mejorar estos porcentajes nace este Proyecto Fin de Carrera. A lo largo de este Proyecto se ha experimentado con diferentes parametrizaciones basadas en diferentes modelos.

Se ha tomado como referencia las parametrizaciones, ya clásicas, plp y mfcc. Partiendo de los procesos que estas realizan, pero utilizando otros modelos auditivos, se han implementado una serie de nuevas parametrizaciones. Estos modelos han sido los de Lyon, Seneff y el modelo ERB.

En una segunda parte del Proyecto se ha tratado de introducir el Procesado Morfológico en esta étapa de parametrización.

Además de la capacidad de estos modelos para imitar el oído humano y de la posibilidad de incorporar el Procesado Morfológico se ha jugado con el contexto como elemento fundamental en las parametrizaciones.

Por tanto, gracias a los resultados obtenidos, se ha puesto de manifiesto la viabilidad de utilizar estos nuevos modelos auditivos y la posibilidad de realizar un procesado morfológico como parte de la parametrización.

La Memoria del Proyecto Fin de Carrera ha sido estructurada y redactada procurando facilitar la tarea de su lectura. Suponiendo que accederán a esta Memoria personas con distintos niveles de conocimientos e inquietudes sobre el tema tratado se ha dividido el documento en cuatro grandes bloques.

De esta manera se podrá realizar la lectura únicamente de aquella parte, o partes, en las que se tenga interés.

- Primer bloque: comprende los capítulos 1, 2 y 3. En este primer bloque se explica la base teórica en la que se fundamenta este Proyecto (Modelado Auditivo, Parametrizaciones, Procesado Morfológico, etc). También se hace un repaso sobre el estado del arte en el Reconocimiento Automático del Habla.

- Segundo bloque: el capítulo 4. En este capítulo se detalla el Entorno de Experimentación: la base de datos utilizada y el sistema de pruebas.
- Tercer bloque: formado por los capítulos 5 y 6. Se explican los distintos experimentos que han sido llevados a cabo. También se analizan los resultados obtenidos y se sacan conclusiones de este análisis.
- Cuarto bloque: este último bloque lo conforman una serie de Anexos sobre las herramientas utilizadas en el Proyecto así como el presupuesto del mismo. Al final se ha inlcuido una recopilación de las siglas y abreviaturas que aparecen en la Memoria.

Índice General

Introducción III

1. Fundamentos del Reconocimiento Automático del Habla 1
1.1. Extracción Clásica de Parámetros . . . 3
1.1.1. Ténicas de Estimación Espectral . . . 4
Bancos de filtros . . . 4
Cepstrum Real . . . 4
Cepstrum Complejo . . . 5
Espectro LPC . . . 7
1.1.2. Parametrizaciones Habituales del Habla . . . 9
Parámetros perceptuales estacionario: PLP y MFCC . 9
Parámetros Característicos Dinámicos . . . 12
Parametrizaciones Robustas frente al Ruido: CMS y RASTA . . . 14
1.2. Métodos de Modelado Acústico . . . 15
1.2.1. Modelos Ocultos de Markov . . . 15
Reconocimiento . . . 16
Entrenamiento . . . 18
Inconvenientes . . . 18
1.2.2. Reconocimiento híbrido: MLP/HMM . . . 19

2. Modelado del Sistema Auditivo 21
2.1. Producción del habla . . . 21
2.2. Sistema Auditivo: Psicoacústica, Morfología y Fisiología . . . 22
2.2.1. Psicoacústica . . . 22
Umbrales Auditivos . . . 22
2.2.2. El sistema receptor auditivo . . . 28
Morfología del Sistema Receptor . . . 28
Fisiología del Sistema Receptor . . . 29
2.2.3. El nervio auditivo . . . 39
Enmascaramiento de sonidos y resolución en frecuencia 39
Efectos Biaurales . . . 43
Sensaciones Tonales . . . 44
2.3. Modelos del Sistema Auditivo . . . 46

2.3.1. Modelo de Lyon 46
2.3.2. Modelo ERB 48
2.3.3. Modelo de Seneff 48
2.4. Conclusiones 49

3. Procesado Morfológico **50**
3.1. Morfología matemática 50
3.2. Transformada Slope o de la Pendiente 51
3.3. Algebra plus-prod vs Algebra max-plus 52
3.4. Transformada de Legendre-Fenchel 53
3.5. Transformada de Cramer 53
3.5.1. Comparación con Cepstrum 54

4. Entorno de Experimentación **55**
4.1. La Base de Datos 55
4.2. El Sistema de Pruebas 56

5. Experimentos y Resultados **58**
5.1. Experimento de Referencia 59
5.2. Nuevos Modelos Auditivos 61
5.2.1. Modelo de Lyon 61
Experimento 1 62
Experimento 2 63
Experimento 3 64
5.2.2. Modelo ERB 65
Experimento 1 66
Experimento 2 66
5.2.3. Modelo de Seneff 67
5.2.4. Resultados 68
5.3. Transformada de Cramer 70
5.3.1. Modelo de Lyon 70
5.3.2. Modelo ERB 71
5.3.3. Modelo de Seneff 72
5.3.4. Resultados 72

6. Análisis de los Resultados **74**
6.1. Conclusiones Finales 81
6.2. Líneas Futuras 81

A. Instalación del Sistema de Pruebas **82**
A.1. SPRACHcore 82
A.2. Quicknet3 83
A.3. Dpwelib 83

B. El "AuditoryToolbox" 85
B.1. LyonPassiveEar 85
B.2. MakeERBFilters 86
B.3. SeneffEar 87
B.4. Sobre el código 87

C. Presupuesto del Proyecto 88

D. Abreviaturas y Siglas 90

Lista de Figuras

1.1. Esquema básico de un Sistema de ASR, extraído de [10] 2
1.2. Análisis cepstral por deconvolución[16]. 6
1.3. Comparación de LPC y PLP [16] 13
1.4. Esquema básico de HMM[8] 16

2.1. Ejemplo de umbrales de audición humanos [13] 23
2.2. Variación del jnd en función de la intensidad absoluta para un tono puro y para ruido de ancho de banda limitado (arriba, línea continua); Y en función del ancho de banda (abajo) [13] . 24
2.3. Umbrales diferenciales para cambios de frecuencia [13] 25
2.4. Niveles de igual volumen para ruido en función de su ancho de banda [13] 27
2.5. Tono percibido de un tono puro en función de su frecuencia [13] 27
2.6. Representación del oído [9] 28
2.7. Curvas de Sintonía de una fibra nerviosa auditiva del gato [13] 31
2.8. Amplitud relativa y fase de la vibración a lo largo de la cóclea [13] 34
2.9. Enmascaramiento anterior y posterior [13] 41
2.10. Variación del ancho de banda crítica en función de la frecuencia central [13] 42
2.11. Esquema básico del modelo de Lyon [23] 47
2.12. Gammatone [21] 48

5.1. Señal de Referencia 62
5.2. Bandas de la Señal de Referencia con el Modelo Lyon1 62
5.3. Cocleograma de la Señal de Referencia con el Modelo Lyon1 . 63
5.4. Bandas de la Señal de Referencia con el Modelo Lyon2 64
5.5. Cocleograma de la Señal de Referencia con el Modelo Lyon2 . 64
5.6. Bandas de la Señal de Referencia con el Modelo Lyon3 65
5.7. Cocleograma de la Señal de Referencia con el Modelo Lyon3 . 65
5.8. Bandas de la Señal de Referencia con el Modelo ERB1 66
5.9. Cocleograma de la Señal de Referencia con el Modelo ERB1 . 66
5.10. Bandas de la Señal de Referencia con el Modelo ERB2 67
5.11. Cocleograma de la Señal de Referencia con el Modelo ERB2 . 67

5.12. Bandas de la Señal de Referencia con el Modelo Seneff 68
5.13. Cocleograma de la Señal de Referencia con el Modelo Seneff . 68

6.1. Mejora de los experimentos desajustados, contexto 5 75
6.2. Mejora de los experimentos desajustados, contexto 7 75
6.3. Mejora de los experimentos clean-clean, contexto 5 76
6.4. Mejora de los experimentos clean-clean, contexto 7 76
6.5. Mejora de los experimentos noisy-noisy, contexto 5 77
6.6. Mejora de los experimentos noisy-noisy, contexto 7 77
6.7. Influencia del contexto en la mejora, experimentos sin Transformada de Fenchel 78
6.8. Influencia del contexto en la mejora, experimentos con Transformada de Fenchel 79
6.9. Mejora de los experimentos con el modelo Lyon 79
6.10. Mejora de los experimentos con el modelo ERB 80
6.11. Mejora de los experimentos con el modelo Seneff 80

Lista de Tablas

3.1. Comparación TMS vs TLS[12]. 51
3.2. Definición de los algebras 'max-plus' y 'min-plus'[12] 52

5.1. Tasa de Error, por palabra. Experimentos de Referencia. . . . 60
5.2. Porcentaje de Mejora y Significación Estadística para MFCC. 60
5.3. Tasa de Error, por palabra. Nuevos Modelos. 69
5.4. Porcentaje de mejora, respecto a PLP. Nuevos Modelos. 69
5.5. Tasa de Error, por palabra. Parametrizaciones con Cramer. . . 73
5.6. Porcentaje de mejora, respecto a PLP. Parametrizaciones con Cramer. 73

C.1. Fases del Proyecto 88
C.2. Costes de material 89
C.3. Presupuesto 89

Capítulo 1

Fundamentos del Reconocimiento Automático del Habla

Un Sistema de Reconocimiento Automático de Habla (RAH) pretende obtener una transcripción de la información que, en forma de onda sonora, está recibiendo como entrada. Es un sistema complejo, compuesto por diferentes bloques. Un esquema típico formado por cinco subsistemas [16] puede observarse en la figura 1.1.

Observamos en el esquema un primer bloque encargado de recoger la señal de voz. Lo formarían una serie de micrófonos, amplificadores, filtros, etc. Del diseño e implementación de este primer bloque dependerá el ancho de banda, la relación señal a ruido y las reverberaciones, entre otras características, de la señal con la que trabajaremos en el siguiente bloque.

El segundo bloque, sobre el cual profundizaremos en la sección 1.1 de este trabajo, forma el subsistema de extracción de parámetros característicos de la señal de voz. Este subsistema deberá proveernos de una representación de la señal de voz que sea sensible al contenido fonético y no a las variaciones acústicas de la señal. Esto es, que distintas repeticiones de una misma palabra, en distintas situaciones, nos dé una misma salida. Para conseguir esto se podrán utilizar diversas técnicas: típicamente se enventanará la señal de voz con ventanas de longitud constante y se trabajará con estas porciones de señal.

En el siguiente bloque se obtendrá una medida de similitud entre la muestra obtenida en el bloque anterior y unas muestras de referencia. Esta medida de similitud podrá ser tanto un valor determinista como una medida de probabilidad. A lo largo de este proyecto trataremos siempre el caso probabilístico.

Optimizar estas medidas es la tarea del cuarto bloque. Por lo general se emplearan con este fin algoritmos de programación dinámica, un ejemplo de

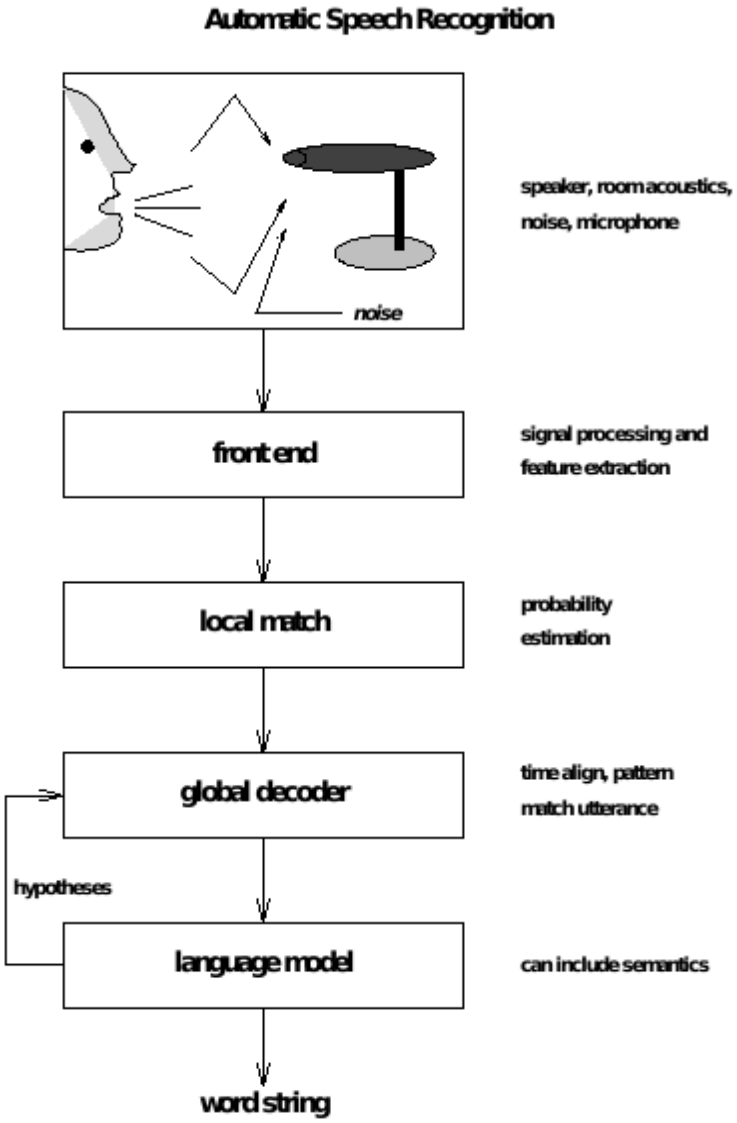

Figura 1.1: Esquema básico de un Sistema de ASR, extraído de [10]

los cuales es el algoritmo de Viterbi, muy utilizado en el caso que se trata en la sección 1.2, los HMM o modelos ocultos de Markov.

En los sistemas de reconocimiento de habla complejos, no será suficiente con la información que cada una de las palabras, o unidad lingüística con la que se trabaje, aporte de forma aislada. Será necesario integrar información sobre el contexto, la sintaxis y la semántica de las palabras. El aporte de esta información es tarea del último de los bloques que aparecen en nuestro esquema.

Ahora que conocemos la estructura básica general de un sistema automático de reconocimiento del habla, podemos profundizar en algunos de sus puntos.

1.1. Extracción Clásica de Parámetros

Como acabamos de ver, la extracción de parámetros a partir de la señal de voz es uno de los pasos fundamentales en el Reconocimiento Automático de Habla

El objetivo principal, a la hora de decidir qué y cuántos parámetros extraemos, será poder diferenciar muestras que pertenezcan a clases distintas. Esto es, buscaremos aquellos parámetros que sean similares para muestras de una misma clase y diferentes para muestras pertenecientes a clases distintas. Tratando además de conseguir una representación que sea estable para distintos ejemplos del mismo sonido, desechando diferencias en el hablante o el entorno[16].

Para conseguir este objetivo se examinarán ventanas de la señal de voz. Esto es, pequeñas porciones (entre 10 y 30 ms). Y se extraerá de ellas algún tipo de coeficiente, típicamente espectral, que condense la información que nos es útil en vectores numéricos (normalmente de longitudes entre 5 y 40 coeficientes)[16].

Aunque en principio podríamos diseñar un vector de parámetros que no tuviese en cuenta el funcionamiento real del oído humano, utilizaremos los conocimientos que de él tenemos. Gracias a estos conocimientos sobre fisiología del sistema audítivo, y el aparato fonador, podemos desarrollar un sistema de extracción de parámetros acústico-articulatorios de la señal de voz robusto y genérico.

En el capítulo 2 se aporta información muy útil sobre estos aspectos:

- Producción del habla, sección 2.1
- Sistema auditivo, sección 2.2

Sería aconsejable leer en primer lugar estas secciones si no se poseen conocimientos previos sobre el tema.

1.1.1. Ténicas de Estimación Espectral

Las técnicas que describimos tratarán de estimar alguna forma del espectro a corto plazo de la señal. O lo que es igual, la envolvente de la Transformada de Fourier de la señal de voz enventanada, una vez que se ha minimizado el efecto del tono en la señal de voz. Ya que en aplicaciones de reconocimiento de voz es irrelevante.

Estas aproximacionesnos serán útiles para realizar la parametrización de la señal de voz.

En concreto, las tres aproximaciones que describiremos serán:

- la de **bancos de filtros**: tratará de modelar dos aspectos del sistema auditivo. La sintonización y el aumento del ancho de banda segun aumentan las frecuencias características de los haces del nervio auditivo.
- la del **"Cepstrum"**: realiza un análisis de la señal de voz teniendo en cuenta cómo ha sido ésta formada. Este tema es tratado en la sección 2.1.
- la del **espectro "LPC"** desarrollado a partir del conocimiento sobre el moldeado que de la señal acústica se hace en el tracto vocal.

Bancos de filtros

Un banco de filtros, como su nombre indica, es una secuencia de filtros paso-banda que permite trabajar de forma independiente con distintas porciones del espectro de la señal de voz. Esto es especialmente adecuado ya que no existe la misma cantidad de información útil en todas las bandas de la señal de habla. En el momento de diseñar estos filtros habrá que tener en cuenta que el oído humano utiliza una ancho de banda mayor cuanto más alta es la frecuencia a la que trabaja. Y que las frecuencias fuera del rango [300 Hz, 3000 Hz] son de menor importancia para el oído[2].

Cepstrum Real

La estructura básica de un modelo de producción de voz puede identificarse con la salida de un sistema de resonadores cuya entrada es una excitación. La convolución de la excitación con la respuesta al impulso del sistema resonador produciría nuestro modelo de la señal de voz[16].

Por tanto, parece lógico realizar un análisis de la señal de voz en el que se separe la fuente (excitación) del filtro (resonadores). A este proceso se le denomina *deconvolución* y al resultado un *análisis "cepstral"* de la señal de voz[16].

Si asumimos que la señal consiste en una secuencia de tiempo discreta, el espectro de esta señal será la transformada Z evaluada en la circunferencia unidad. Considerando esta premisa, y definiendo X como el espectro de la señal observada, E el espectro de la excitación y V el del resonador del tracto vocal, tenemos[16]:

$$|X(\omega)| = |E(\omega)||V(\omega)| \tag{1.1}$$

y tomando logaritmos

$$\lg|X(\omega)| = \lg|E(\omega)| + \lg|V(\omega)| \tag{1.2}$$

Si estas dos componentes varían de forma diferente según ω, como ocurre en el caso de sonidos sonoros, podremos separar de forma sencilla ambas componentes, mediante algún tipo de filtrado. Como estamos en frecuencia y no en tiempo la nomenclatura de procesado de señal cambia: tendremos "liftrado" (liftro) en lugar de "filtrado" (filtro), respuesta en "quefrency" en lugar de respuesta en frecuencia y "cepstrum" en vez de espectro (del lat. "spectrum"), que será la Transformada de Fourier (o TZ) de $\lg|X(\omega)|$. El cepstrum se calcula tomando la TF, o la TZ, inversa de $\lg|X(\omega)|$:

$$c(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \lg|X(\omega)|e^{j\omega n}d\omega \tag{1.3}$$

siendo $c(n)$ los coeficientes "cepstrales".

Para comprender mejor cómo podemos deconvolucionar la señal de voz utilizando un análisis cepstral nos ayudaremos de la figura 1.2[16]:

Cepstrum Complejo

El análisis cepstral explicado hasta este momento sólo hace uso de la magnitud de las señales, dejando a un lado su fase. Por este motivo, es conocido como "cepstrum real". Es posible definir un 'Cepstrum Complejo' que nos permita trabajar también con fases.

En este caso, expresamos la transformada Z de la señal observada mediante[16]:

$$X(z) = \frac{A\prod_{k=1}^{M_i}(1-a_k z^{-1})\prod_{k=1}^{M_o}(1-b_k z)}{\prod_{k=1}^{N}(1-c_k z^{-1})} \tag{1.4}$$

Y siendo $\hat{x}(n)$ el "cepstrum complejo" de X, nos queda:

$$\hat{x}(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \lg X(\omega)e^{j\omega n}d\omega \tag{1.5}$$

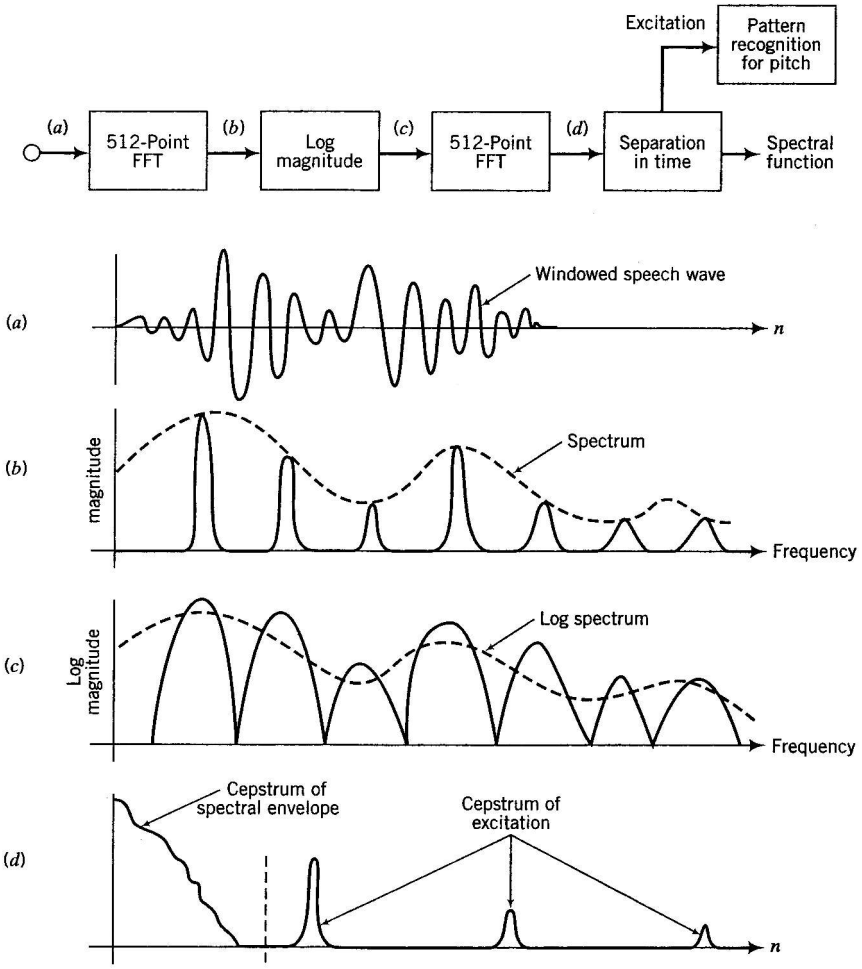

Figura 1.2: Análisis cepstral por deconvolución[16].

Que puede ser evaluado:

$$\hat{x}(n) = \begin{cases} \lg A, & n = 0 \\ \sum_{k=1}^{N} \frac{(c_k)^n}{n} - \sum_{k=1}^{M_i} \frac{(a_k)^n}{n}, & n > 0 \\ \sum_{k=1}^{M_o} \frac{(b_k)^{-n}}{n}, & n < 0 \end{cases} \tag{1.6}$$

De lo cual concluimos que:

- Si $\hat{x}(n) = 0$, para $n \geq 0$, debe corresponder a un filtro todo ceros (es decir, FIR), con todos los ceros fuera de la circunferencia unidad[16].
- Si $\hat{x}(n) = 0$, para $n < 0$, debe corresponder con un filtro con todos los polos y los ceros dentro de la circunferencia unidad. Lo que se define como un "filtro de fase mínima"[16].

Existen evidencias físicas de que las vibraciones que ocurren dentro de la cóclea, y por consiguiente las curvas de sincronización del nervio auditivo, pueden representarse como filtros de fase mínima[16]. Por tanto esta es otra posible aplicación del análisis cepstral.

Aunque en ocasiones será necesario recurrir al cepstrum complejo, no siempre será así, ya que el cálculo de la fase añade una importante complejidad al problema y no está claro que ventajas extra aporta.

Por último, cabe mencionar que un numero limitado (10-14) de coeficientes cepstrales suele ser suficiente para representar el espectro y permitir diferenciar unos sonidos de otros[16].

Espectro LPC

Como hemos comentado anteriormente en el tracto vocal se establecen resonancias. La cantidad de estas resonancias puede ser predicha de forma razonable por modelos de tubos. Las frecuencias de resonacia de dicho tracto vocal son los formantes, fundamentales para la inteligibilidad de la señal de voz.

Si asumimos que para modelar el tracto vocal sólo es necesario una aproximación capaz de representar un número suficente de resonancias, cada formante puede ser representado mediante una función de transferencia de un solo polo con la forma[16]:

$$H_i(z) = \frac{1}{1 - b_i z^{-1} - c_i z^{-2}} \tag{1.7}$$

Asumiendo además un ancho de banda de $5kHz$, necesitaremos cinco resonadores en cascada para representar los cinco formantes que se esperan de media en un humano adulto. Se requieren también uno o dos polos más (posiblemente reales) para representar el espectro. Por tanto, una vocal podría ser representada, mediante seis secciones de forma directa de la siguiente manera[16]:

$$H(z) = \frac{1}{1 - \sum_{j=1}^{P} a_j z^{-j}} \tag{1.8}$$

Siendo P el doble del número de las secciones de segundo orden necesarias de las anteriormente descritas.

Teniendo en cuenta que el espectro a corto plazo de una señal de voz puede representarse mediante un filtro que puede ser especificado por $P = 2(BW + 1)$ coeficientes, con BW en kHz, la respuesta en tiempo discreto $y(n)$ a una excitación $x(n)$ puede describirse como[16]:

$$y(n) = x(n) + \sum_{j=1}^{P} a_j y(n - j) \tag{1.9}$$

Considerando esta fórmula, trataremos de predecir la señal de voz mediante una suma con pesos de los valores previos de la señal:

$$\tilde{y}(n) = \sum_{j=1}^{P} a_j y(n-j) \tag{1.10}$$

Una vez llegados a este punto se nos plantean tres preguntas:

- Qué criterio de error debemos minimizar para la predicción.
- Cuál es el mejor número de coeficientes a usar.
- Cómo calculamos estos coeficientes

El siguiente pseudo-algoritmo constituye una respuesta global típica a estos problemas[16]:

1. Minimización del error cuadrático medio. Sea el error de predicción, $e(n) = y(n) - \tilde{y}(n)$, que queremos minimizar. Minimizar la media de este error cuadrático $E\{e^2(n)\}$ supondría también minimizar la siguiente ecuación:

   $$D = \int_{-\pi}^{\pi} \frac{|Y(\omega)|^2}{|H(\omega)|^2} \frac{d\omega}{2\pi} \tag{1.11}$$

   que conduce a estimar la envolvente del espectro de la señal.

   Además esta minimización también disminuye los efectos del ruido aditivo. Sin embargo, no es necesariamente el criterio ideal, ya que se da importancia a porciones del espectro con grandes amplitudes que pueden no sernos útiles en nuestra aplicación. Este inconveniente puede disminuirse realizando un *pre-énfasis* de la señal.

2. Respecto al número de coeficientes óptimos es lógico pensar que cuantos más coeficientes más detallada será la representación. Y es cierto, pero un mayor nivel de detalle puede hacer la estimación más sensible al error.

   Por tanto el orden del modelo dependerá de la aplicación a diseñar. En nuestro caso, RAH, el orden del modelo será muy cercano a la regla dada anteriormente, es decir, P.

3. Para calcular los coeficientes, a, debemos seguir el siguiente razonamiento:

   $$e(n) = y(n) - \sum_{j=1}^{P} a_j y(n-j) \tag{1.12}$$

   $$D = \sum_{n=0}^{N-1} e^2(n) = \sum_{n=0}^{N-1} [y(n) - \sum_{j=1}^{P} a_j y(n-j)]^2 \tag{1.13}$$

Si tomamos derivadas parciales nos queda:

$$\sum_{j=1}^{P} a_j \phi(i,j) = \phi(i,0) \text{ para } i = 1, 2, ..., P \tag{1.14}$$

donde $\phi(i,j)$ es una correlación entre versiones desplazadas de la señal de voz.

Los coeficientes que producen la mejor aproximación de $\tilde{y}(n)$ a $y(n)$ en el sentido de menor error cuadráico medio como la descrita arriba son los llamados *coeficientes de predicción linear* (ing. "Linear Predictive Coefficients", LPC). La diferencia entre la estimación y la señal ($e(n) = y(n) - \tilde{y}(n)$) se denomina el *residuo de predicción LPC*[16]. Este error es la parte de la señal que no es predecible a partir de sus valores previos, en este caso, la excitación.

En la práctica, los LPC no son una buena representación para muchas aplicaciones debido a que[16]:

- Los coeficientes polinómicos son muy sensibles a la precisión numérica.
- La estabilidad del filtro resultante no esta garantizada.
- Los coeficientes no son ortogonales ni estan normalizados.

Por estos motivos los LPC suelen ser transformados generalmente en alguna otra forma de representación, como los "Rout pairs", coeficientes de reflexión, el Cepstrum, etc.

1.1.2. Parametrizaciones Habituales del Habla

En las ultimas décadas distintas variantes de los bancos de filtros, la LPC y el cepstrum han sido usadas en la extracción de parámetros para RAH. Recientemente la mayoría de los sistemas ha convergido al uso de un vector de parametros cepstrales obtenido a partir de un banco de filtros[16].

Parámetros perceptuales estacionario: PLP y MFCC

A continuación nos centraremos en dos de estas variantes, los MFCC (ing. "Mel Frequency Cepstral Coefficients") y los coeficients PLP (ing., "Perceptual Linear Prediction"), que son muy similares.

Ambas técnicas tratan de estimar los parámetros mediante procesos que simulen la forma en que funciona el oído humano y cómo éste percibe sonidos con distintas componentes frecuenciales. Modelan el proceso de activación de

las Células Ciliadas Internas por los movimientos vibratorios de la membrana basilar[18].

Los pasos básicos en ambos análisis son:

1. **Calcular una estimación de la potencia espectral para la ventana analizada**. Esto es, se enventana la señal (por ejemplo con una ventana de Hamming) y se calcula la FFT y el cuadrado del modulo de la misma[16].

2. **Integrar la potencia espectral en las bandas críticas de las respuestas de los filtros**. Se utilizan distintas clases de filtros, pero todos ellos estan basados en una escala de frecuencia que es aproximadamente lineal por debajo de $1kHz$ y aproximadamente logarítmica por encima de ese punto, conocida como la *escala MEL*, basada en la percepción del tono y se utiliza con bancos de filtros para calcular los MFCC[1]. Estos filtros estan basados en experimentos con humanos y existen distintas aproximaciones y modelos[16].

   - MFCC: la integración se realiza con una ventana triangular sobre el logaritmo de la potencia del espectro.
   - PLP: se utilizan filtros de pendiente trapezoidal aplicados aproximadamente en intervalos de 1 "Bark". El eje "Bark" deriva del eje frecuencial usando una función de distorsión no lineal (ing., "warping"[2]):

$$\Omega(w) = 6 \ln \left\{ \frac{\omega}{1200\pi} + \left[\left(\frac{\omega}{1200\pi} \right)^2 + 1 \right]^{0.5} \right\} \tag{1.15}$$

   Su efecto neto es reducir la sensibilidad frecuencial de la estimación espectral original, en particular en altas frecuencias[16].

3. **Pre-enfatizar el espectro para aproximarlo a la desigual sensibilidad del oído humano a las distintas frecuencias**[16].

   - MFCC: en la mayoría de análisis mel-cepstral este paso se realiza antes del análisis del espectro original y un efecto importante es que elimina la componente contínua de la señal de voz.
   - PLP: este paso es implementado como una ponderación explícita de los elementos del espectro de bandas críticas.

4. **Comprimir la amplitud espectral**. El efecto de este paso es reducir las variaciones de amplitud de las resonancias espectrales[16].

[1] Más información sobre la escala mel en la sección 2.2.1
[2] Más información sobre bark en la sección 2.2.3

- MFCC: aplicamos el logaritmo después de la integración.
- PLP: se toma la raíz cúbica en lugar de el logaritmo.

5. **Realizar la** DFT^{-1}. Es un paso crítico en ambos métodos[16].

   - MFCC: en este paso se generan los coeficientes.
   - PLP: dado que no se ha realizado el logaritmo, los resultados son más similares a coeficientes de correlación (el espectro aquí esta comprimido).

   Dado que los valores de la potencia son reales, sólo los componentes del coseno de la DFT^{-1} necesitan ser calculados.

6. **Suavizado espectral**. Aunque el espectro de bandas críticas suprima algún detalle, otro nivel de integración es util en la reducción de los efectos de fuentes no-lingüisticas de varianza en la señal de voz[16].

   - MFCC: este paso es realizado por truncamiento cepstral: típicamente sólo se conservan los 12 ó 14 menores componentes de los 20 ó más filtrados. Con ello se consigue una representación del espectro suavizado.
   - PLP: un modelo autorregresivo (derivado de la solución de ecuaciones lineales construidas a partir de las autocorrelaciones del paso previo) es utilizado para suavizar el espectro comprimido de bandas críticas. Como en el caso del espectro LPC, el espectro suavizado resultante realiza una medida mejor en los picos que en los valles. Muchos investigadores encuentran que esta aproximación permite mayor robustez al ruido e independencia del hablante que el truncamiento cepstral.

7. **Ortogonalizar la representación**[16].

   - MFCC: no es necesaria ya que los elementos del cepstrum truncado ya son ortogonales
   - PLP: los coeficientes autoregresivos son convertidos a variables cepstrales.

8. **Liftrado**. Es un paso adicional que suele realizarse en el que los parámetros cepstrales son multiplicados por una función sencilla como:

$$n^{\alpha} \tag{1.16}$$

   Siendo n el indice 'cepstral' y $\alpha \in [0, 1]$. El propósito de esta función es modificar las distancias que podrían calcularse con estas características para que sean más o menos sensibles a las amplitudes de los picos resonantes del espectro[16].

Cuando $\alpha = 1$ se dice que el cepstrum está ponderado con un índice y el liftro resultante tiene el efecto de igualar las varianzas de los distintos coeficientes cepstrales[16].

Las características de los MFCC ylos coeficientes PLP son extremadamente similares. Ambos tienen baja resolución en altas frecuencias, indicativo de métodos basados en bancos de filtros y proveen salidas ortogonales, típico del análisis ceptral. Y ambos proporcionan una representación correspondiente a un espectro a corto plazo alisado que ha sido comprimido y cuantificado de la manera en que lo haría un oído humano[16]. La principal diferencia entre ambos procedimientos radica en la naturaleza del espectro suavizado (basado en coeficientes cepstrales o en coeficientes LPC).

Además es importante resaltar que mientras MFCC hace uso de la escala mel, PLP utiliza la escala Bark. Si bien es cierto que ambas escalas son escalas de bandas críticas en las cuales los filtros estan distribuidos a lo largo del eje frecuencial de forma lineal hasta los 1000 Hz y de forma logarítmica por encima de esta frecuencia[18].

Algunos investigadores que han experimentado mezclando ambos procedimientos, han encontrado ventajoso usar 'PLP' pero con la ventana de integración triangular[16].

Dado que los parámetros PLP están basados en los LPC resulta interesante comparar ambos análisis: la figura 1.3 muestra en paralelo sendos diagramas de bloques de uno y otro proceso. El análisis LP de la figura respresenta los LP-cepstra, que se obtienen con una recursión una vez han sido calculados los parámetros LP.

Hay que hacer notar que el paso de autocorrelación de LPC está expandido a su equivalente en el dominio de la frecuencia ($DFT^{-1}|\cdot|^2$) en esta figura. En esencia el pre-énfasis hecho en el análisis LPC es equivalente al filtrado isofónico (ing. "equal-loudness filtering") en PLP. La principal diferencia radica en la computación del espectro en bandas críticas comprimido en PLP, que no se realiza en LPC.

Consecuentemente, el cálculo PLP puede verse tanto como un análisis 'mel-cepstral' con suavizado espectral del tipo LP; o como un análisis LPC para una versión implícita de la voz que ha sido modelada de acuerdo con las propiedades acústicas. En la práctica se ha observado que con esta técnica se registra una mayor independencia del hablante[16].

Parámetros Característicos Dinámicos

La voz es un proceso no estacionario pero se asume que es cuasi-estacionario ya que para un periodo corto de tiempo los estadísticos de la voz no difieren demasiado. Los análisis MFCC y PLP estiman el espectro localmente como

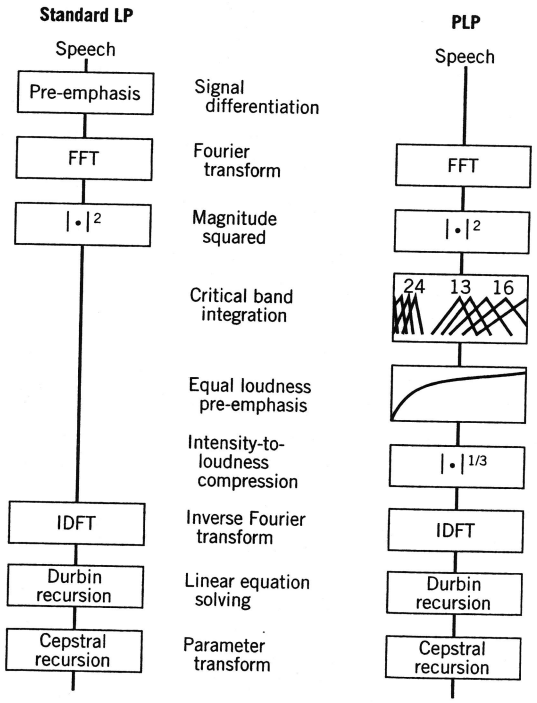

Figura 1.3: Comparación de LPC y PLP [16]

un proceso cuasi-estacionario. Sin embargo, una de las características de la voz es su comportamiento dinámico. Debido a esto, numerosos investigadores utilizan estimaciones de las variaciones temporales de los "cepstrum" o del espectro a corto plazo.

Una de las medidas más comunes es la llamada *delta del cepstrum*, que es una aproximación por mínimos cuadrados de la pendiente local[16], y como tal es una estimación suavizada de la derivada local de la diferencia entre tramas vecinas del cepstrum. La expresamos como:

$$\Delta C_i(n) = \frac{\sum_{k=-N}^{N} kC_i(n+k)}{\sum_{k=-N}^{N} k^2} \tag{1.17}$$

Cada flujo de valores delta cepstral se calcula correlando los correspondientes flujos de valores cepstrales con una recta de pendiente unidad.

La segunda derivada, la *delta de la delta del cepstrum*, o aceleración, es también muy útil, y corresponde a una correlación similar pero en este caso con una función parabólica[16].

Muchos sistemas de reconocimiento[16] han incorporado características como estas que tratan de modelar y recalcar los aspectos dinámicos del espectro de la señal de voz. Sin embargo las características así obtenidas son insuficientes para caracterizar totalmente a la señal, por lo que en la mayoría

de los casos los sistemas que incorporan características delta lo hacen como un complemento a las medidas estáticas, ya sean MFCC o PLP.

Parametrizaciones Robustas frente al Ruido: CMS y RASTA

Los experimentos[16] han demostrado que las parametrizaciones de voz explicadas en epígrafes anteriores tienen una sensibilidad excesiva al ruido presente en la señal de entrada. Para combatir dicha influencia se han diseñado otras parametrizaciones que ahora pasamos a describir.

Substracción Cepstral de la Media, CMS Recordemos que podemos definir el espectro observado a corto plazo de una señal procesada por un filtro lineal e invariante en el Tiempo como:

$$X(\omega, t) = S(\omega, t)H(\omega, t) \tag{1.18}$$

o en potencias logarítmicas:

$$\lg |X(\omega, t)|^2 = \lg |S(\omega, t)|^2 + \lg |H(\omega, t)|^2 \tag{1.19}$$

Si las componentes (S y H) tienen propiedades distintas a lo largo del tiempo pueden ser separadas de forma sencilla. Si H es constante a lo largo del tiempo y si los componentes constantes de S no nos son útiles, es posible estimar la componente constante de la suma calculando la media del espectro logarítmico. De forma alternativa, puede calcularse la transformada de Fourier de los componentes y restar las medias en ese dominio. Esta operación es estándar en muchos sistemas RAH, y es conocida como *substracción cepstral de la media*, CMS, (ing. "Cepstral Mean Subtraction")[16].

Gracias a este procedimiento podemos eliminar una perturbación en la señal, ya que aunque en tiempo estaría convolucionada con la señal original, ahora esta sumada.

De un modo más general, CMS puede verse como un ejemplo específico de un filtrado en el dominio del tiempo de coeficientes cepstrales o en el espectro de log-potencia[16].

Modificación RASTA RASTA-PLP es s una modificación del análisis PLP como una aproximación en linea para lograr robustez contra perturbaciones convolucionales[16].

Dado que el oído humano es menos sensible a los cambios lentos en el sonido, en esta aproximación se ha tratado de suprimir o disminuir las componentes de variaciones lentas. Para ello se ha reemplazado el espectro a corto plazo en bandas críticas convencional del PLP por una estimación espectral,

en la cual cada canal frecuencial es filtrado paso-banda mediante un filtro con un cero a frecuencia cero. Mediante esta operación cualquier constante o variación lenta (bajas frecuencias) es eliminada. El nuevo espectro es menos sensible a las bajas frecuencias en el espectro a corto plazo[3].

RASTA ("RelATive SpecTrAl") puede verse como una versión a corto plazo de CMS, o como una reintegración de los coeficientes 'delta', y, aunque en comparación con PLP mejore la independencia del error convolucional, hay que tener en cuenta el efecto de las condiciones iniciales para su cálculo[16].

1.2. Métodos de Modelado Acústico

Como comentamos al inicio del capítulo, tras la fase de extracción de parámetros característicos de la señal de voz pasábamos a comparar estos parámetros, mediante alguna medida de similitud, con ciertos valores previamente almacenados, que se denominan *modelos acústicos*.

Aunque la medida de similitud utilizada podría estar basada en valores determinísticos, y de hecho existen sistemas RAH que funcionan con ellos, nos centraremos en la utilización de valores probabilísticos.

En concreto, el modelo que utilizaremos para comparar los parámetros que nos llegan de la señal de voz con aquellos que representarían las distintas palabras, u otras unidades del lenguaje, serán los Modelos Ocultos de Markov o HMM (ing. "Hidden Markov Model") en singular

1.2.1. Modelos Ocultos de Markov

Un modelo de Markov se define como una máquina de estados estocásticos finitos. Cambia de estado, siguiendo una distribución probabilística, una vez en cada instante de tiempo. Y por cada instante de tiempo en que se produce un cambio de estado se genera una nueva salida, teniendo también en cuenta ciertas densidades de probabilidad, que en este contexto se denomina *observación*[19].

Por tanto se definen dos procesos estocásticos concurrentes:

1. **La secuencia (oculta) de estados internos**.

   Sea $Q = \{q_1, ..., q_k, ..., q_K\}$ el conjunto finito de todos los estados posibles. Un HMM específico, M_i, se representa mediante un Automata de Estados Estocasticos Finitos, SFSA (ing. "Stochastic Finite State Automaton"), con L_i estados $S_i = \{s_1, ..., s_l, ..., s_{L_i}\}$, $s_l \in Q$, siguiendo una topología específica de transición entre estados[10].

[3]Para conocer más en profundidad esta parametrización puede consultarse [17].

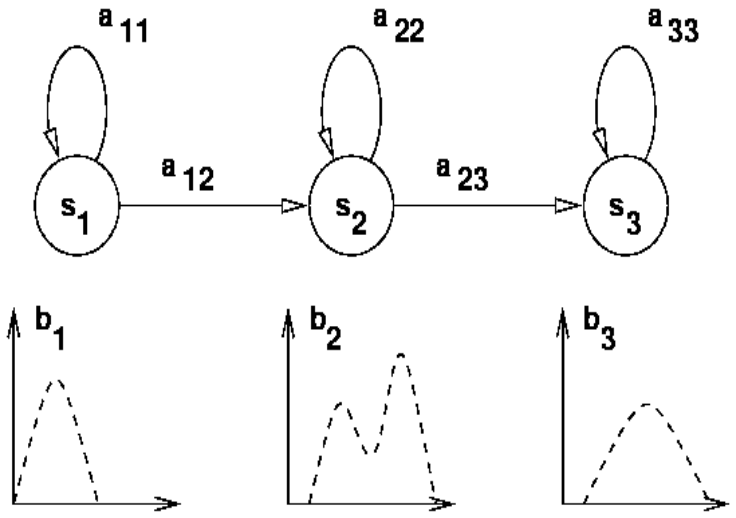

Figura 1.4: Esquema básico de HMM[8]

2. **Los procesos de salida (observaciones).**

   Un HMM modela la secuencia de vectores de parámetros $X = \{x_1, ..., x_n, ..., x_N\}$ como un proceso estacionario en el cual cada segmento estará asociado con un estado HMM específico[10].

Esto es, usando un modelo M, una secuencia $X = \{x_1, ..., x_N\}$ es generada mediante una sucesión de estados $S = \{s_1, ...s_L\}$, siendo $L \leq N$ y existiendo transiciones instantáneas entre estados. Un esquema de un HMM sencillo puede observarse en la Figura 1.4.

En el caso límite existiría un HMM para cada posible expresión a reconocer. Lo cual sería inviable, porque no tendríamos el suficiente número de secuencias de observaciones para entrenar dichos modelos. Por tanto adoptamos un esquema jerárquico para reducir, notablemente, el número de posibles modelos.

En primer lugar se modela una oración como una secuencia de palabras, y a continuación cada palabra es modelada mediante unidades lingüisticas más pequeñas, que bien pueden ser sílabas, semi-sílabas o fonemas. Siendo esta última la categoría lingüistica más comúnmente utilizada[19].

Reconocimiento

Dado un conjunto de modelos entrenados para reconocer determinadas unidades y una secuencia de observaciones X que se supone proveniente de dichos modelos, en el paso de Reconocimiento, un sistema debe ser capaz de encontrar el modelo, o modelos, de la unidad lingüistica empleada, que con

mayor, probabilidad ha sido la que ha generado la secuencia de observaciones a reconocer. Puesto en forma de ecuación, tratamos de encontrar el modelo M que máximiza $P(M|X)$. Usando el teorema de Bayes podemos escribir:

$$P(M|X,\Theta,\Theta^*) = \frac{P(X|M,\Theta)P(M|\Theta^*)}{P(X|\Theta)} \tag{1.20}$$

Siendo Θ y Θ^* los conjuntos de parámetros de los modelos acústico, $P(X|M,\Theta)$, y lingüistico $P(M|\Theta^*)$.

$P(X|M)$ y $P(M)$ son estimados durante la fase de reconocimiento por subsistemas cuyos parámetros han sido entrenados con muestras diferentes. Por este motivo, los dos conjuntos de parámetros, Θ (en el modelo acústico) y Θ^* (en el modelo lingüistico), se consideran independientes. Debido a esto y asumiendo que la estimación de $P(X)$ es independiente de Θ (lo cual es falso en entrenamiento) la ecuación 1.2.1 permite la formulación de un entrenamiento del modelo acústico como un problema de Estimación de Máxima Verosimilitud, (ing. "Maximum Likelihood Estimation", MLE). Nótese que $P(X|\Theta)$ es constante para X dada y no entra en la maximización[10].

La verosimilitud acústica se calcula entonces expandiéndola en todos los posibles caminos en M que pudieran generar X:

$$p(X|M,\Theta) = \sum_{\forall S^j} p(X,S^j|M,\Theta) \tag{1.21}$$

Esta verosimilitud puede aproximarse sustituyendo el sumatorio por el máximo, lo que se conoce como de la aproximación de Viterbi:

$$p^*(X|M,\Theta) = \max_{\forall S^j} p(X,S^j|M,\Theta) \tag{1.22}$$

y puede ser empleada en reconocimiento sin sufrir importantes pérdidas.

En la fase de reconocimiento de una secuencia X desconocida, tendremos que escoger el mejor modelo M_j que maximice $P(M_j|X,\Theta,\Theta^*)$:

$$\begin{aligned} j &= \arg\max_{\forall i} P(M_i|X,\Theta,\Theta^*) \\ &\approx \arg\max_{\forall i} p(X|M_i,\Theta)P(M_i|\Theta^*) \end{aligned} \tag{1.23}$$

Esta busqueda suele ser resuelta mediante el algoritmo de Viterbi, que es un caso particular de programación dinámica, (ing., "Dynamic Programming", DP). Puede encontrarse una explicación de este algoritmo en [19].

Entrenamiento

Durante la fase de entrenamiento debemos determinar Θ y Θ^* que maximicen $P(M_j|X_j, \Theta, \Theta^*)$ para todas las secuencias de entrenamiento X_j[10]:

$$\Theta, \Theta^* = \arg\max_{\Theta,\Theta^*} \prod_{j=1}^{J} P(M_j|X_j, \Theta, \Theta^*) \tag{1.24}$$

Sin embargo en los sistemas HMM acústicos suponemos que podemos maximizar por separado

$$\Theta = \arg\max_{\Theta} \prod_{j=1}^{J} P(X_j|M_j, \Theta) \tag{1.25}$$

Estos parámetros Θ, durante la fase de reconocimiento, permitirán calcular las probabilidades de emision $P(x_n|q_k, \Theta)$, que multiplicadas y asumiendo independencia estadística producen la $P(X|M)$ buscada.

Existen algoritmos de entrenamiento eficientes para los parámetros Θ. El más común es un caso particular del de maximización de la esperanza, EM en sus siglas inglesas, "Expectation-Maximization", que es el algoritmo "adelante-atrás" (ing. "forward-backward"). Para consultar este algoritmo referirse a [19].

Inconvenientes

Hemos podido comprobar que HMM provee una forma eficiente de trabajar. Sin embargo, para poder llegar a las conclusiones que aquí se han enunciado ha sido necesario realizar importantes asunciones, que son la principal causa de debilidad de los HMMs. Algunas de estas hipótesis que han sido asumidas se enumeran a continuación[10].

1. El algoritmo de entrenamiento esta diseñado siguiendo un criterio MLE en lugar de máximo a posteriori, MAP, lo que provoca una mala discriminación final.

2. Se asume que las secuencias de estados de los HMMs son cadenas de Markov de primer orden.

3. Se elige una topología de densidades de probabilidad a priori, por lo general Gaussianas o mezcla de ellas.

4. Al intentar mejorar la eficiencia de los algoritmos se suponen independientes ciertos procesos que no lo son.

1.2.2. Reconocimiento híbrido: MLP/HMM

Como acabamos de ver, HMM es una herramienta muy útil para modelar Sistemas de Reconocimineto de Habla. Sin embargo han de tenerse en cuenta sus limitaciones.

Vamos a explicar a continuación un sistema híbrido que combina HMM y redes neuronales artificiales (ing. "ANN"). Estos sistemas híbridos han sido diseñados para solventar las deficiencias de los HMM.

Las redes neuronales pueden diseñarse para realizar de forma muy satisfactoria numerosas tareas, clasificación, regresión, etc., e incluso pueden utilizarse en sistemas RAH sin estar asociadas a un HMM, en cuyo caso sólo se obtienen tasas de reconocimiento aceptables en tareas de reconocimiento de palabras aisladas[10].

En el caso que vamos a detallar la red neuronal empleada sera un MLP, perceptrón multicapa, (ing., "Multi-Layer Perceptron") y calculará una probabilidad similar a las probabilidades de emisión, $P(x_n \mid q_k)$, utilizadas en los HMMs[4], de forma que puedan ser fácilmente integradas en el modelo HMM.

Podemos entrenar nuestro MLP para producir $P(q_k \mid x_n)$ si cada salida del MLP esta asociada con un estado específico del HMM[10]. Estas probabilidades pueden ser convertidas en probabilidades de emisión si aplicamos Bayes:

$$\frac{P(q_k|x_n,\Theta)}{P(q_k)} = \frac{p(x_n|q_k,\Theta)}{p(x_n|\Theta)} \tag{1.26}$$

siendo es este caso Θ el conjunto de parámetos del MLP.

La utilización de ANN para realizar esta subtarea soluciona en gran medida todos los incovenientes citados en 1.2.1.

Además, distintos experimentos realizados con sistemas híbridos MLP/HMM señalan importantes mejoras respecto a los sistemas clásicos HMM[10]:

1. Los sistemas híbridos relativamente sencillos han resultado muy eficientes (desde el punto de vista de consumo de CPU y requerimientos de tiempo de memoria), a la par que precisos.

2. Sistemas más complejos son entrenados de forma sencilla con muy buenos resultados en tareas complejas de reconocimiento contínuo de voz.

3. Utilizando un mismo número de parámetros se obtienen mejores resultados con un sistema híbrido que con un HMM convencional.

[4]Para un conocimiento preciso sobre el diseño y entrenamiento de este tipo de MLP referirse a [11] ó [10]

4. La inclusión de información sobre el contexto en los sistemas MLP/HMM es sencilla.

5. Los HMM funcionan muy bien para la parametrización clásica MFCC. MFCC se ajusta a las asunciones de modelado que fija HMM, además han sido configurados de forma conjunta. Esto dificulta la posibilidad de experimentar con nuevas parametrizaciones, que es el objetivo de este Proyecto. Los sistemas híbridos no presentan este inconveniente.

Para terminar, hay que citar que también se ha investigado en sistemas que combinen ambos procedimientos, MLP/HMM y HMM[10].

Debido a todos estos motivos se ha optado por utilizar este tipo de modelos híbridos MLP/HMM en la realización de todos los experimentos de este Proyecto Final de Carrera.

Capítulo 2

Modelado del Sistema Auditivo

Como se mencionó en la sección 1.1 resulta de gran importancia conocer la manera en que la voz es producida y cómo el sistema auditivo procesa los sonidos que recibe.

Por tanto, a lo largo de este capítulo trataremos de explicar estos dos importantes procesos. En la última sección de este capítulo se introducen tres modelos auditivos basados en los principios expuestos en las dos primeras. Estos modelos serán los utilizados para las experimentos del Proyecto.

2.1. Producción del habla

El mecanismo de producción del habla en el ser humano es complejo por lo que en esta sección sólo se intentará dar una ligera idea de aquellos aspectos que puedan ser relevantes para la comprensión del resto de los apartados de la memoria. [1]

El aparato fonador puede modelarse como un tubo resonante de geometría variable y controlable, formado por la laringe, la faringe y los tractos paralelos vocal y nasal, que comienza en la glotis y termina en los labios y los orificios de la nariz[16].

La configuración y la posición específica que tomen los distintos elementos anatómicos a lo largo del tubo resonante, también llamados *articuladores*, caracterizarán la señal de voz.

La estructura básica del modelo de producción de voz puede ser identificado con una excitación, producida en las cuerdas vocales, que entra en un sistema resonador variable formado por la laringe, la faringe y los tractos vocal y nasal, en el que el paso del aire se ve obstruido por los articuladores. Ambos componentes, excitación y resonación, juegan un importante papel

[1] Para ampliar información sobre este tema refiéranse a [16].

en la producción de la voz, de funciones distintas. La excitación es responsable de una cantidad menor de carácterísticas, como la naturaleza sorda o sonora del sonido, mientras que la resonación produce la otra gran mayoría de parámetros que nos permitirán clasificar los sonidos[16].

2.2. Sistema Auditivo: Psicoacústica, Morfología y Fisiología

El oído humano, junto con el cerebro, es capaz de realizar un procesado de la señal sonora muy útil desde el punto de vista de la comprensión. Por ejemplo, somos capaces de filtrar ciertos tipos de ruidos y diferenciar variaciones en la frecuencia e intensidad del sonido. Además se ha demostrado que el nervio auditivo tiene fibras especializadas en las distintas frecuencias de la señal del espectro de audición humana[16].

El conocimiento que tenemos sobre el funcionamiento del neocortex y el nervio auditivo, lugares en donde se realizan importantes funciones relacionadas con la audición, es limitado. Como una nota curiosa sobre el funcionamiento del nervio auditivo que nos pueda inspirar a la hora de diseñar sistemas RAH, comentaremos la existencia de caminos de retorno o retroalimentación entre el cerebro y el oído interno[16].

2.2.1. Psicoacústica

La psicoacústica es la ciencia que describe la percepción de sonidos por parte de los humanos, u otros animales. A continuación se enumeran y comentan brevemente ciertas caracterísitcas relevantes sobre la audición animal y humana conocidas gracias a la investigación psicoacústica.

Umbrales Auditivos

- **El Umbral Absoluto**
  Definimos el umbral absoluto como la intensidad sonora mínima detectable. Este umbral mínimo es dependiente de la frecuencia.
  El umbral se considera infinito para frecuencias inferiores a 20Hz, y para las superiores a 20 kHz. El rango audible humano comprende 10 octavas[13]. Podemos ver un audiograma en la figura 2.1.

- **Umbrales diferenciales**

  - **Umbrales Diferenciales de Intensidad**
    Son los cambios mínimos del nivel de intensidad(δL) que pueden

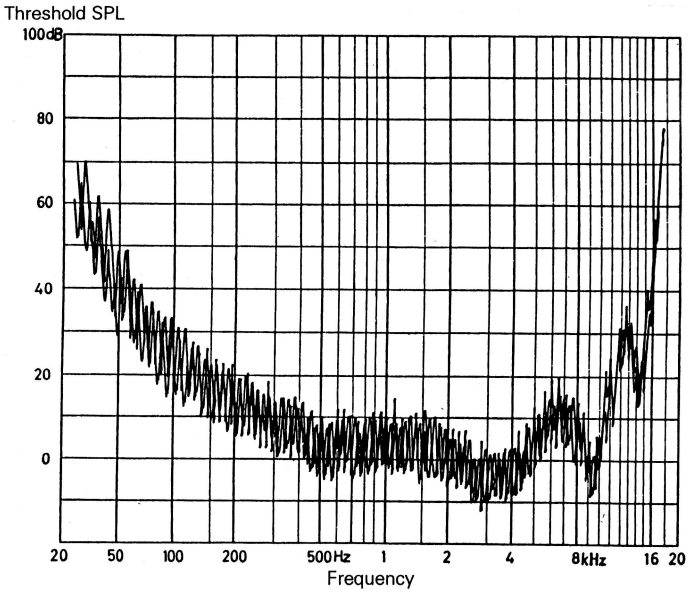

Figura 2.1: Ejemplo de umbrales de audición humanos [13]

ser percibidos (ing. "jnd: just-noticeable difference")[13]. Las ideas que a continuación se exponen pueden quedar más claras a la vista de la figura 2.2.

- Tonos Puros
  Si consideramos un ratio de modulacion cercano a 4 Hz podemos obtener los siguientes resultados: Un jnd de 2 dB para intensidades de 15 dB, un jnd de 1 dB para intensidades de 30 dB, y un jnd de 0.5 dB para intensidades de 60 dB. El jnd es independiente de la frecuencia auditiva[13].
- Ruido Blanco
  Al contrario de lo que ocurria en el caso anterior el jnd disminuye siempre que aumentemos la intensidad sonora[13].
- Ruido Paso-Banda
  El jnd aumenta al disminuir el ancho de banda del ruido[13].

- **Umbrales Diferenciales para cambios de frecuencias**
  Buscamos la variación de frecuencia mínima detectable δf, que dependerá de:
  1. La velocidad a la que la modulación se produce (f_{mod}) y la forma de la modulación.
  2. La audiofrecuencia f.
  3. El nivel de intensidad.
  4. La duración del estímulo.

  Podremos representar los umbrales diferenciales tanto mostrando δf como función de f, figura 2.3 (arriba), o mostrando $\delta f/f$ como

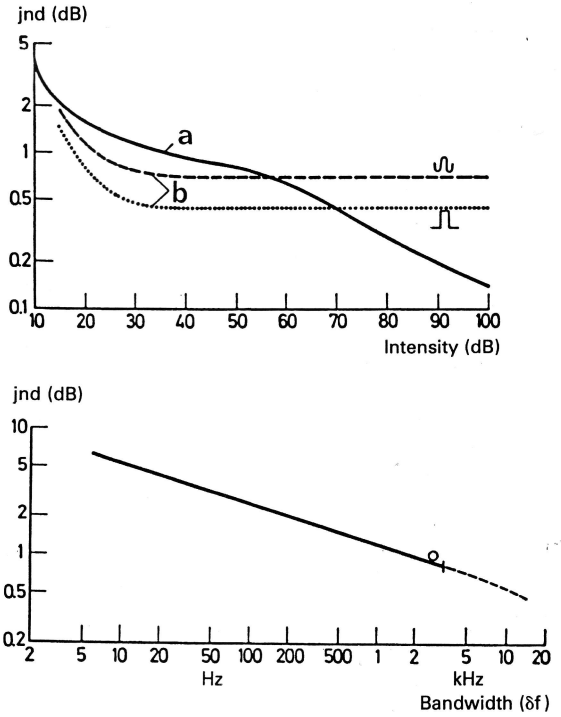

Figura 2.2: Variación del jnd en función de la intensidad absoluta para un tono puro y para ruido de ancho de banda limitado (arriba, línea continua); Y en función del ancho de banda (abajo) [13]

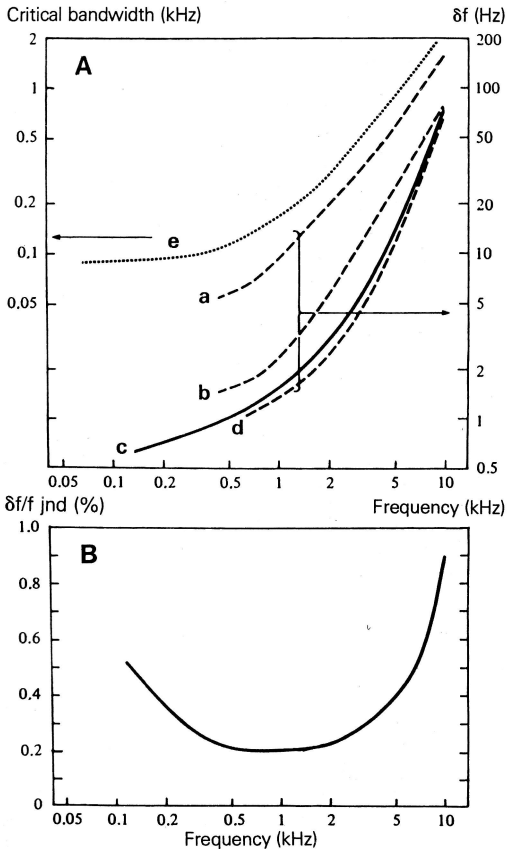

Figura 2.3: Umbrales diferenciales para cambios de frecuencia [13]

una función de f, figura 2.3 (abajo). En cualquier caso deberemos de tener siempre en cuenta el nivel de intensidad utilizado.

Debemos resaltar la relativa constancia de $\delta f/f$ para un cierto rango grande de frecuencias. Además el jnd no varía mucho (para una frecuencia fija) al variar la intensidad. Tendiendo eso sí a ser menor cuanto más aumentemos la intensidad[13].
Por último, refiriéndonos a la duración del impulso y considerando una duración normal de 200 ms, el umbral jnd δf mejora apreciablemente (disminuye) al aumentar la duración d.

- **Audición de tono puros por encima del Umbral**
  En este apartado trataremos la percepción subjetiva de la intensidad acústica, volumen sonoro, y de la frecuencia o tono percibido.

  - **Escalas de Volumen e Intensidad Sonora**
    La percepción subjetiva de la intensidad también varía con la frecuencia[16].

- La Escala de Niveles Iguales de Volumen. Fonos
  Distintos niveles objetivos de intensidad son percibidos como iguales. Aquellos niveles de intensidad (a 1 kHz) que son percibidos como un sonido de n dB, se les considera $nfonos$[13].
- La Escala de Niveles Diferentes de Volumen. Sonos
  En este caso tratamos de buscar estimaciones diferenciales de volumen. Esto es aquellos niveles de intensidad que se perciben como el doble o la mitad que otro. En este caso utilizamos la unidad sono. La subjetiva apreciación de doble intensidad (a 1 kHz) corresponde con un aumento objetivo de 10 dB[13]. Se establece un valor de un sono a un sonido a 1 kHz producido a 40 fonos.

$$N_{sonos} = 2^{0.1(L_N - 40)} \tag{2.1}$$

  Donde L_N es el nivel de igual intensidad acÃºstica medida en fonos.

$$\log_{10} N = 0.03(L_N - 40) \tag{2.2}$$

$$N_{sonos} = (I/I_0)^{0.3}/16 = (p/p_0)^{0.6}/16 \tag{2.3}$$

  En estas formulas podemos observar una relación logarítmica entre el estímulo físico y su percepción subjetiva. De hecho si representamos en una gráfica log/log sonos frente a nivel de presión observaremos una relaciíon lineal entre 40 dB y 80 dB.
- El volumen de las señales de ruido
  Debemos también considerar el volumen de las señales de ruido y su relación con el ancho de banda del ruido. Los resultados de experimentos [13] ponen de manifiesto que para todos los niveles de intensidad el efecto del ruido es independiente del ancho de banda, siempre que este sea menor que cierto valor llamado *banda crítica* [ver también la sección 2.2.3]. En la figura 2.4 pueden observarse distintas intensidades de ruido con una banda crítica de 160 Hz.
- Combinaciones de sonidos de frecuencias idénticas o muy similares
  Consideremos dos sonidos sinusoidales con idénticos niveles de intensidad y frecuencias f_1 y f_2.
  - Si las dos frecuencias son identicas:
    y los sonidos estan en fase, se produce un refuerzo del sonido. Si por el contrario estan en contrafase se produce un efecto contrario.
  - Si las dos frecuencias son muy similares:
    El sonido resultante tiene una amplitud modulada a un ratio $(f_1 - f_2)$. Este fenómeno es conocido como *batido*[13].

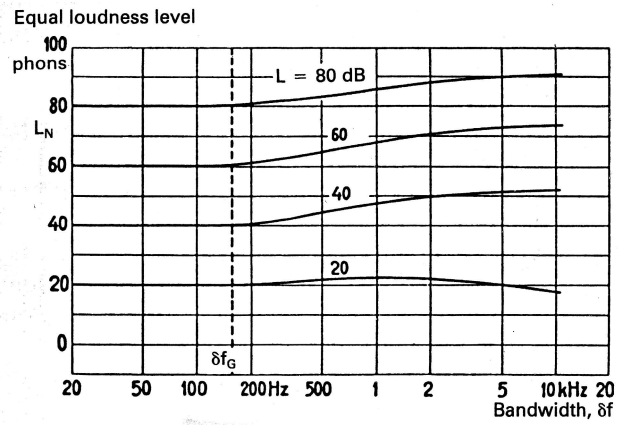

Figura 2.4: Niveles de igual volumen para ruido en función de su ancho de banda [13]

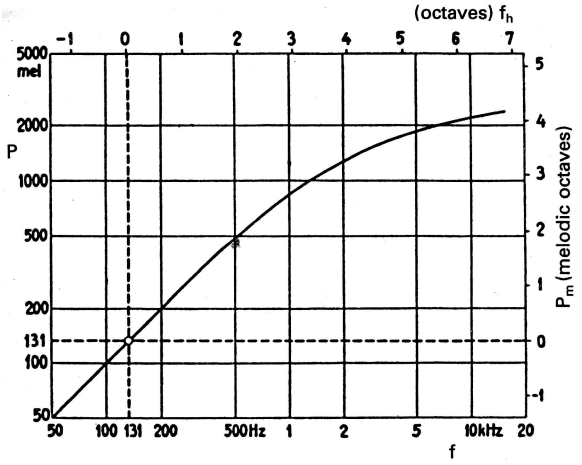

Figura 2.5: Tono percibido de un tono puro en función de su frecuencia [13]

- **Una Escala de tono percibido para tonos puros**
  Pasaremos ahora a considerar la sensación subjetiva de la frecuencia en los sonidos. Utilizaremos una unidad llamada *mel*: 500 mels corresponden al tono percibido de un tono a 500 Hz a 40 dB[13]. Utilizando esta escala para dibujar curvas de tono percibido respecto a la frecuencia, como la de la figura 2.5, observamos que el tono aumenta menos que la frecuencia para frecuencias altas. Por debajo de los 500 Hz, se oberva una relación lineal que pasa a ser más o menos logarítmica para valores superiores.
  - Nivel de Tono Percibido y Nivel de Intensidad
    El tono que percibimos de un sonido variará no solo por su frecuencia, también puede hacerlo en función de su intensidad.

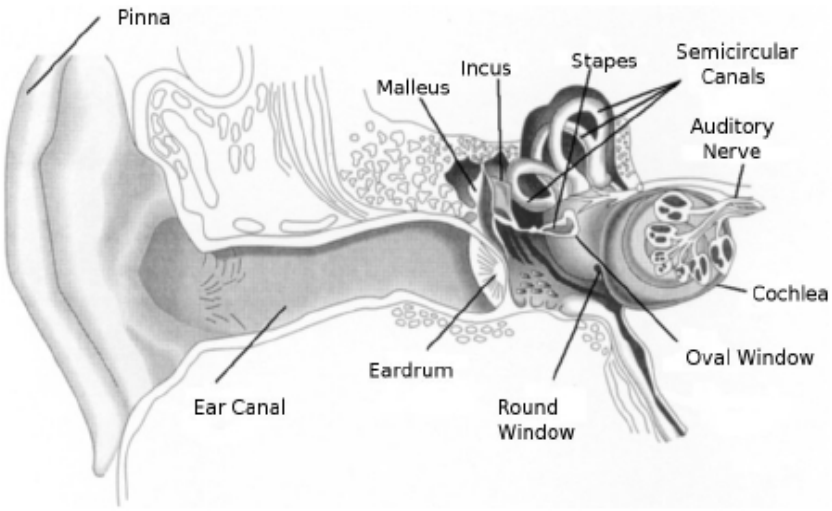

Figura 2.6: Representación del oído [9]

2.2.2. El sistema receptor auditivo

Comencemos tratando de explicar, tanto desde un punto de vista morfológico como fisiológico, el sistema receptor auditivo.

Morfología del Sistema Receptor

El sistema receptor auditivo está formado por tres componentes[20]:

1. El oído externo
2. El oído medio
3. El oído interno

Podemos ver una representación en la figura 2.6.

- **El Oído Externo**
  El oído externo comprende el canal auditivo externo, la concha y la pinna. La pinna en lo seres humanos está esencialemente formada por cartílago recubierto de piel. Esta región cartilaginosa se extiende hasta la concha, la cual forma el tercio exterior del canal auditivo[13].

- **El Oído Medio**
  El oído medio, o cavidad timpánica, se situa entre los oídos interno y externo. Esta comunicado con la faringe a traves de las trompas de

Eustaquio.
Al final del canal auditivo externo se encuentra el tímpano. La cavidad timpánica esta comunicada con el oído interno a través de la ventana oval y la ventana redonda. El tímpano y la ventana oval se encuentran unidos por la cadena osicular, una pequeña estructura osea formada por los huesos: martillo, yunque y estribo[13].

- **El oído Interno**
  El oído interno, alojado en el hueso temporal esta constituido por una serie de cavidades intercomunicadas: la cóclea, el vestíbulo, y los canales semicirculares. Todas ellos forman el laberinto de huesos.

  En la cóclea, el laberinto conecta con el oído medio a través de la ventana oval y su membrana, así como a través de la ventana redonda, que también esta cerrada por su propia membrana.
  En sección transversal la cóclea aparece dividida en dos partes con forma de espiral: la escala vestibular, superior y conectada con el vestíbulo; y la inferior escala timpánica que termina en la ventana redonda. Estas dos divisiones estan conectadas mediante el helicotrema.
  La escala vestibular y la timpánica pertenecen al espacio perilinfático. Entre ellas se encuentra el canal coclear, la escala media, que esta separada de la segunda por la membrana basilar y de la primera por la membrana de Reissner. El canal coclear pertenece al espacio endolinfático.
  La membrana de Reissner esta formada por dos capas de células. Es permeable y parece que ciertos intercambios se producen entre la perilinfa y la endolinfa.
  Dentro de la escala media, y a lo largo de toda la cóclea, se encuentra el órgano de Corti que contiene las células receptoras. Es en este órgano donde la energía vibratoria se transduce en energía electroquímica codificada neuronalmente.

  Existen dos tipos de células receptoras: las células ciliadas externas(en inglés, "outer hair cells" OHC) y las células ciliadas internas(en inglés, "inner hair cells" IHC). Las OHC no descansan sobre la membrana basilar, lo que si hacen las IHC.
  La cóclea se enrolla a través de un eje óseo, la columela, la cual acomoda la entrada de las fibras del nervio coclear[13][20].

Fisiología del Sistema Receptor

El estímulo sonoro primero entra por el oído externo y medio cuya función es asegurar una correcta transmisión de la energía acústica hasta el oído interno, donde se encuentran los mecanismos receptores[16].

- **Transmisión del Sonido**

  - **El Odio Externo**
    La compleja geometría del oído externo ayuda a localizar la procedencia de los sonidos y produce una ganancia en la presión sonora. Esta ganancia es dependiente de la frecuencia del sonido[13].
  - **El oído Medio**
    Las vibraciones acústicas del aire hacen vibrar al tímpano. Este movimiento es transmitido a través de yunque, el martillo y el estribo hasta la ventana oval. Después la energía sonora es transmitida al órgano de Corti. La disposición de la cadena de huesecillos juega un importante papel en la transferencia de la vibración[13].
    El oído Medio como un Transformador de Impedancias
    Si las vibraciones del aire actuaran directamente sobre la ventana oval al menos un 98 % de la energía sería reflejada. Gracias a la cadena de huesecillo un 67 % de la energía es absorbida[13].
    Además proporciona un control automático de ganancia.

- **Recepción del Sonido y Transducción en la Cóclea**

  - **Señales Auditivas en las Neuronas Primarias**
    Trataremos el problema de la correcta transducción y la codificación de información auditiva en impulsos nerviosos.
    En primer lugar consideraremos la excitación de los receptores por la entrada auditiva que reciben, y despues estudiaremos la estructura resultante de las señales auditivas aferentes.
    - Actividad Espontánea:
      En ausencia de cualquier tipo de entrada sonora la mayoría de los receptores generan de forma espontánea descargas en sus aferentes con una tasa muy variable[13].
    - Codificación de Frecuencia y Fase:
      Una importante característica de todas las fibras es la existencia de su propia curva de sintonía. Si la intensidad mínima, el cambio mínimo detectable, para excitar una fibra se muestra en función de la frecuencia la curva resultante típicamente tendrá un mínimo en una frecuencia a la que denominaremos característica, f_c.
      Puede observarse en la figura 2.7 que estas curvas de sintonía son típicamente irregulares. El umbral aumenta más rápidamente con las frecuencias por encima de f_c que con las inferiores. Normalmente, la pendiente de las umbrales ascendentes

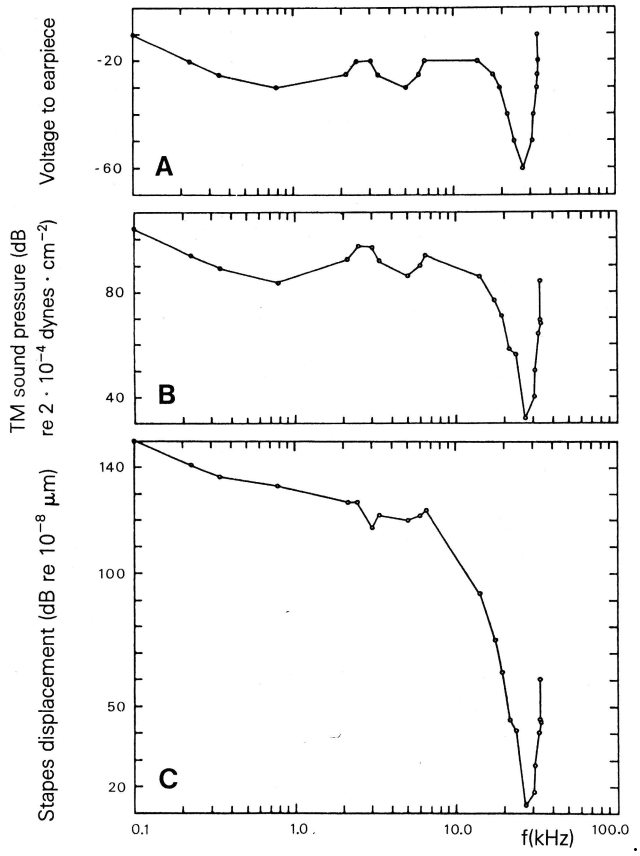

Figura 2.7: Curvas de Sintonía de una fibra nerviosa auditiva del gato [13]

son mayores en aquellas fibras con frecuencias características altas ($f_c > 2kHz$). Por lo que la sintonización es relativamente más pronunciada para las fibras con altas f_c.
Una variedad de parámetros ha sido utilizada para especificar las pendientes de estas curvas de sintonía, y con ello los anchos de banda de los filtros paso-banda que caracterizan el funcionamiento del oído interno. Se utilizan los dB/octava para medir las pendientes cuando $f > f_c$ y $f < f_c$. Se define Q_{10dB} como el ratio adimensional de f_c en la banda de frecuencia δf comprendido entre las frecuencias de la curva que dan una intensidad de umbral 10 dB por encima de f_c. $Q_{10bB} = f_c/\delta f$. También es común referirse al ancho de banda efectivo, el ancho de banda en Hz comprendido entre las frecuencias de la curva de intensidad de umbral 3 dB por encima de f_c.
El umbral mínimo de intensidad detectable a f_c varía en las distintas fibras. Al menos parte de la variación de sensibilidad de la cóclea según la frecuencia es debida a la variación del oído medio con la frecuencia.
Cuando la cóclea es estimulada a bajas frecuencias pero con suficiente intensidad, toda la base de la cóclea, esto es, todos los receptores con los valores de f_c más altos son activados a la vez[13].

- Codificación de la Intensidad:
  Cuando se aumenta un estímulo, a una frecuencia fija, la descarga es también aumentada, en muchos casos de forma sigmoidal[13].

- **Fenómenos Bioeléctricos en la Cóclea**
  Al tratar de descubrir cómo las células ciliadas son excitadas por las ondas sonoras se han descubierto una variedad de fenómenos biológicos.

  - Potenciales Constantes (DC):
    Existen constantes diferencias de potencial entre las distintas partes de la cóclea que son independientes de cualquier tipo de estimulación[13].
  - Potenciales Microfónicos Cocleares:
    Los microfónicos cocleares (CM, "cochlear microphonics") son una actividad eléctrica (potenciales) que precede a la generación del impulso nervioso en el proceso de trasducción de la cóclea al nervio. Los CM siguen perfectamente la amplitud instantánea del estímulo asemejando la salida de un micrófono.
    El logaritmo del CM guarda una relación lineal con la inten-

sidad del estímulo en decibelios.
Los CM son generados por las células ciliadas[13].

- Suma de Potenciales (SP):
Aparte de los CM es posible detectar en la ventana oval o en el canal coclear diferencias de potencial DC que se mantiene durante la duración del estímulo sonoro. La distribución espacial de las Sumas de Potenciales a lo largo de la cóclea presenta un máximo a altas frecuencias hacia la base y a bajas frecuencias hacia el ápice[13].

- **Mecanismos Cocleares**
A continuación veremos todos los mecanismos por los cuales una vibración acústica que llega a la ventana oval termina excitando los receptores cocleares.

  - Datos de Partida e Hipótesis:
  Las distintas frecuencias sonoras afectan a regiones distintas de la cóclea, entre la base y el ápice. Historicamente se han planteado dos hipótesis para justificar este comportamiento.

    1. La Hipótesis de la resonancia.
    Segun esta teoría la membrana basilar comprende una serie de estructuras resonantes, cada una con una frecuencia resonadora. Situandose las frecuencias altas en la base de la cóclea, cerca de la ventana oval, y las bajas frecuencias cerca del ápice.
    El efecto de un sonido es activar únicamente los resonadores sintonizados a las frecuencias correspondientes con las presentes en el sonido. Según esta teoría el oído sería capaz de descomponer un sonido en una serie de Fourier. Esta teoría también implica la existencia de un numero limitado de fibras dedicadas a cada frecuencia.
    Este modelo provoca una serie de dificultades teóricas: En primer lugar es dificil imaginar como un resonador puede ser atacado con suficiente intensidad pero sin que esta intensidad afecte al resto. Segundo, para ser altamente selectivo en frecuencia el resonador debe ser fuertemente amortiguado. En otras palabras, la audición debería ser pobre para sonidos con variaciones en frecuencia muy rápidas y/o amplitud. Tercero, los cambios de fase en los que un resonador actua sólo comprenden entre $-\pi/2$ y $+\pi/2$[13].
    Por último, para realizar esta hipótesis no se tuvieron en cuenta ciertos efectos del sistema auditivo como las sumas

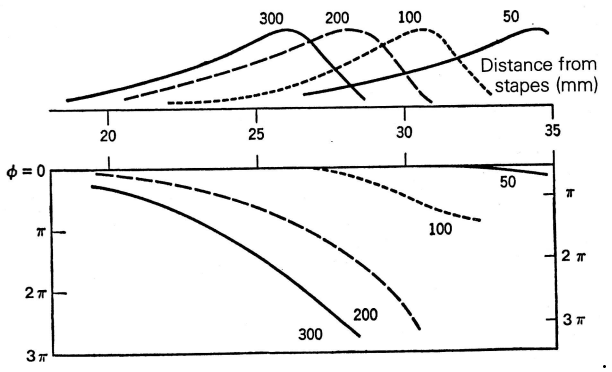

Figura 2.8: Amplitud relativa y fase de la vibración a lo largo de la cóclea [13]

bitonales[2].

2. La Hipótesis de la onda viajera.
   Esta hipótesis se basa en la existencia de una onda viajera que se desplaza a lo largo de la cóclea, desde la base hasta el ápice. La amplitud de esta onda aumentaría inicialemnte para después disminuir.
   Debido a la incompresibilidad de los fluidos del oído interno una variación de presión generada por el estribo provoca un desplazamineto de la endolinfa y un desplazamiento compensatorio de la ventana redonda. Este desplazamiento causa una deformación de la región basal de la cóclea.
   Esta deformación se propaga como una onda transversal desde la base y hacia el ápice. La amplitud de esta onda varía durante su recorrido aumentando en primer lugar hasta un cierto valor y después disminuyendo rápidamente hasta llegar al ápice muy pequeña o casi inapreciable.
   La velocidad de propagación disminuye continua y uniformemente durante todo su recorido.
   Sonidos de distintas frecuencias producen ondas que alcanzan el máximo en distintos puntos. Para altas frecuencias el máximo se encuentra cerca de la base, y para las frecuencias bajas este máximo se alcanza mucho más lentamente y más cerca del ápice. Otra forma complementaria de representar las propiedades de la onda viajera es medir la vibración en un punto concreto de la cóclea en función de la frecuencia, para una amplitud constante. Al representar estos resultados, ver figura 2.8, se observa

[2]Las sumas bitonales son tratadas en la sección 2.2.3

como se alcanza un máximo a cierta frecuencia característica f_c, esta f_c es la frecuencia característica de la fibra en la que se realizan las medidas[13].

- Estudios Más Detallados y Otros Datos

1. Propiedades Mecánicas de Distintos Componentes Cocleares[13]:
   - La membrana basilar es isotrópica en sus propiedades.
   - La membrana basilar no esta normalmente bajo presión.
   - La membrana basilar esta construida de forma que la dirección de propagación de la onda viajera es siempre del punto de base a ápice.
2. Presiones Intracocleares: del Movimiento del Estribo al de la Membrana Basilar.
   Al considerar la hipótesis de la onda viajera existe el problema de cómo los movimientos del estribo se traducen en movimientos perpendiculares de la membrana basilar.
   Respecto a las diferencias de presión entre la escala vestibular y la escala timpánica:
   - En la escala vestibular la presión aumenta rápidamente, a 10 dB/octava hasta que la frecuencia alcanza 1 kHz, después disminuye más lentamente, 5 dB/octava. La fase de la onda de presión varía progresivamente de los $+90^o$ inciales hasta -225^o[13].
   - En la escala timpánica la presión apenas varía al menos hasta 0.5 kHz. A altas frecuencias existe un aumento ligero. la fase es incialmente cero y ocasionalmente alcanza 180^o a 10 kHz[13].

   Estos efectos juegan un rol esencial en las dinámicas cocleares, donde todas las presiones en el tímpano de sonidos por encima de 40 Hz crean diferencias de presión entre la escala vestibular y la timpánica, que son el tipo de presiones transversales necesarias para generar las deformaciones de la membrana basilar a lo largo de la cóclea[13].
3. La Onda Viajera: Amplitud y Fase.
   Como comentamos anteriormente existe el problema de comprensión sobre cómo la onda viajera, un conjunto de oscilaciones transversales que se propagan longitudinalmente con una velocidad finita a lo largo de toda la longitud coclear, pueden ser generadas por movimientos del estribo. No sólo es necesario inventar ecuaciones para representar el fenómeno sino también asegurar la completa

relación entre las dos variables: $a(t)$,la amplitud de la vibración longitudinal del estribo; y $b(t)$, la amplitud de la vibración transversal de la membrana basilar[13]. Las investigaciones han tratado de determinar:

a) El valor de b_m, la amplitud máxima de la vibración a lo largo de la membrana basilar.

b) El valor de b_m en un cierto punto de la membrana basilar, como función de la frecuencia.

c) La fase Φ de la vibración con respecto a los movimientos del estribo y su variación con la frecuencia.

- ◇ Criterio Mecánico para hacer Vibrar la Membrana Basilar: La amplitud del movimiento de la membrana basilar no es proporcional a la amplitud del desplazamiento del estribo sino a su velocidad[13].
- ◇ Movimiento de un Punto de la Membrana Basilar en Función de la Frecuencia: La membrana basilar se asemeja en su funcionamiento a una serie de filtros paso banda[13].
- ◇ Movimiento de un Punto de la Membrana Basilar en Función de la Intensidad del Estímulo: Se ha observado una excelente linealidad entre la intensidad del sonido incidente y la amplitud de la vibración de la membrana basilar en un punto dado, para una frecuencia fija, hasta 120 dB. En la proximidad de f_c se puede observar una cierta no linealidad para intensidades alrededor de 70 a 90 dB[13].
- ◇ Diferencias de Fase entre las Vibraciones del Estribo y la Membrana Basilar: Existen cambios de fase a lo largo de la cóclea entre la vibración del estribo y la del punto en cuestión, en una cierta variedad de frecuencias.
  Junto a la ventana redonda la fase del movimiento de la membrana basilar va $\pi/2$ por delante del movimiento del estribo pero esta en fase con la velocidad del estribo y con la presión sonora.
  Si mantenemos constante la amplitud de las vibraciones del estribo, la fase varía en un punto dado de la cóclea en función de la frecuencia entre 0 y 3π.
  Entre la base y el ápice la fase cambia más rapidamente a frecuencias altas, pero apenas varía a la muy baja frecuencia de 50 Hz.
  Considerando el comportamiento de los puntos de máxima amplitud de la vibración de la membrana basilar a distintas frecuencias, es notable que cerca de este má-

ximo el desfase en referencia al movimiento del estribo varía entre π y 3π[13].

- ◇ Modelar la Cóclea: Modelos de la membrana basilar pueden sernos útiles para entender los mecanismos cocleares. Atendiendo a los movimientos de la membrana basilar comprobamos como este es longitudinal a la altura de la ventana oval (tangencial a la ventana) y rápidamente adquiere un componente normal a la membrana. Este segundo componente es debido a las diferencias de presión entre las dos escalas e incialmente sufre un desfase de $\pi/2$ con respecto al desplazamiento del estribo y consecuentemente en fase con su velocidad. El resultado es que, en teoría, cada punto de la membrana basilar traza un movimiento elíptico. El componente tangencial desciende monótonamente entre la base y el ápice, mientras que la componente normal aumenta lentamente a lo largo de la cóclea hasta un punto de desplazamiento máximo y después desciende rápidamente como la amplitud de la envolvente de la onda viajera[13].

4. Un Nomograma para Distribuciones Frecuenciales.
   Es posible trazar una función, f frente a $\Phi(l)$, siendo f la frecuencia y l la distancia a lo largo de la membrana basilar del punto de máxima amplitud de vibración correspondiente a esa frecuencia. De estas curvas puede reportarse una fuerte relación lineal entre el logaritmo de f y la distancia l[13].
5. No Linealidades en los Mecanismos Cocleares.
   - ◇ Microfónicos Cocleares
     - *a*) Armónicos Aurales Subjetivos: Usar una estimulación por tonos-puros, sin componentes armónicos, de intensidad moderada genera un CM sinusoidal a la frecuencia del sonido incidente. Pero a partir de cierto nivel de intensidad alto aparece una distorsión en el espectro del CM a las frecuencia de los armónicos 2,3, ... Otra importante observación al respecto es que la amplitud máxima del segundo armónico de f_c esta localizada al mismo nivel de la espiral cóclear que el armónico fundamental[13].
     - *b*) Combinación de Tonos: Al usar dos tonos de frecuencias distintas y aumentar la intensidad de uno de ellos a partir de cierto nivel de intensidad aparece un componente extra en CM. Este componente corresponde con un tono diferencial de primer orden. De igual for-

ma aparecen otros tonos combinatorios[13].

c) Supresión bitonal: igual que en los casos anteriores.

◇ Señales Aferentes. Otra manera de estudiar los mecanismos cocleares es registrar las descargas en una única fibra aislada del nervio coclear. Utilizando esta técnica se han observado los siguientes dos ejemplos de interacciones entre sonidos:

a) Supresión bitonal

b) Tono diferencial cúbico

Estos procesos también son considerados creados por el nervio auditivo. En la sección 2.2.3 se aporta más información sobre el tema.

- **Transducción en la Cóclea**
  La manera en la cual un estímulo acústico es finalmente transformado en excitación de las terminaciones nerviosas aferentes.

  ○ Procesos de la Membrana que Acompañan la Excitación de los Cilios Celulares
  Se ha sugerido que las células de la papila basilar actuan como un sistema acústico resonante, que puede representarse con un modelo como un resonador eléctrico que contiene elementos de tipo L (inductancias), C (capacitancias) y R (resistencias)[13].

  ◇ Microfónicos Cocleares, Sumatorios de Potenciales y Potenciales Receptores
  El estímulo acústico provoca un cambio en la resistencia eléctrica transmembranal de las células ciliadas que causa la entrada de una corriente despolarizante.
  Medidas intracelulares han sido realizadas en las regiones basales de la cóclea (la primera parte) cuyas células (IHC y OHC) son claramente mejor estimuladas por las altas frecuencias (3 kHz), pero si el estímulo es suficientemente intenso también responden a bajas frecuencias (300 Hz)[13].

  1. En las IHC, las respuestas intracelulares a bajas frecuencias son oscilatorias a la frecuencia del estímulo. Tienen una respuesta AC como los potenciales microfónicos cocleares. Sin embargo, a altas frecuencias la respuesta no es completamente oscilatoria sino que consta de una respuesta DC despolarizante positiva contínua, que es como un sumatorio de Potenciales (SP)[13].
     El ratio AC/DC (que puede ser comparado con el ratio CM/SP) es alrededor del 20 % a 1kHz y cae a altas frecuencias de 6 a 9 dB/octava.

2. En las OHC, también se registran una serie de despolarizantes/hiperpolarizantes respuestas AC a bajas frecuencias pero esta vez sin ningun signo de efectos rectificantes. No hay componentes DC[13].

2.2.3. El nervio auditivo

Más allá del oído interno encontramos el nervio auditivo. Se han realizado numerosos experimentos para conocer el funcionamiento de este nervio, y se ha podido comprobar que en él se producen numerosas reacciones. Algunas de las cuales, como comentamos anteriormente, son comunes a la membrana basilar[16]:

- Adaptacion: las neuronas son más sensibles a los estímulos cambiantes que a los constantes
- Sintonizacion: los distintos haces del nervio auditivo responden mejor a determinadas frecuencias
- No-linealidades:
  - Saturación: Los distintos haces del nervio auditivo se saturan a distintas intensidades
  - Supresión bi-tonal: La respuesta a un estímulo sonoro puede desaparecer o verse disminuida al aparecer otro tono distinto del primero.
  - Enmascaramiento de un sonido por el ruido.
  - Combinacion de tonos: si una fibra nerviosa es excitada por dos tonos de frecuencias distintas puede aparecer una respuesta a un tono que no existe.

Enmascaramiento de sonidos y resolución en frecuencia

Tendremos en cuenta dos tonos. Un tono de prueba (test) T, con una frecuencia f_T y una intensidad L_T. Por otro lado, tendremos un tono enmascarador M, con una frecuencia y una intensidad f_M y L_M[13]. El enmascaramiento puede ser total o parcial. Un tono aplicado en un oído puede enmascarar otro tono del oído contrario, pero la mayoria de los experimentos se refieren al caso monoaural.
Entre los factores que determinan el enmascaramiento destacan: la intensidad del tono enmascarador, el espectro y las frecuencias relativas de los dos tonos.

- **Enmascaramiento de un tono puro por otro tono puro**
  Consideramos las siguientes conclusiones:

  - El enmascaramiento es más pronunciado cuando las frecuencias f_T y f_M son próximas.
  - Un enmascaramiento es más efectivo si $f_T > f_M$ que si $f_T < f_M$.
  - El efecto enmascarador aumenta linealmente con L_M.

  Cuando la intensidad del tono de prueba es tal que se captan ambos tonos simultaneamente, se aprecian complicados fenómenos. Batidos si f_M y f_T son muy cercanas, o lo son algunos de sus armónicos superiores. En el rango $f_T \neq kf_M$ se aprecian bien una mezcla de ambos tonos, bien su tono diferencial ($|f_M - f_T|$)[13].

- **Enmascaramiento de un Tono Puro por Ruido Blanco**
  Cuando una señal de ruido enmascaradora tiene una cierta intensidad I_R, el umbral de audición de un tono de test L_T (la intensidad mínima necesaria) como una función de la frecuencia del tono de test muestra dos regímenes[13]:

  - Hasta 500 Hz: el umbral no varía.
  - A partir de 500 Hz: el umbral aumenta 4 dB por octava, 10 dB por decima.

  Por el contrario, el enmascaramiento aumenta vertiginosamente como resultado de aumentar la densidad espectral de potencia enmascaradora.
  De todo esto podemos concluir que una señal ruidosa con una densidad espectral de intensidad que sea constante hasta 500 Hz y a partir de ahí disminuye a 10 dB/decada da lugar a un enmascarmiento uniforme e independiente de la frecuencia de tono puro enmascarado[13].

- **Enmascaramiento Simultaneo, Previo y Posterior**
  Introducimos ahora un nuevo parámetro a considerar en los enmascaramientos. El momento en el que un tono de test T es aplicado en relación al sonido enmascarador M.
  Aparte del enmascaramiento simultáneo (T y M a la vez) existe un enmascaramiento residual después de que M desaparezca (posterior) y un empeoramiento del umbral de T justo antes de ser M aplicado (previo)[13].

  Podemos apreciar en la figura 2.9 como:

  - El umbral aumenta rápidamente durante los 10-30 ms anteriores a M.

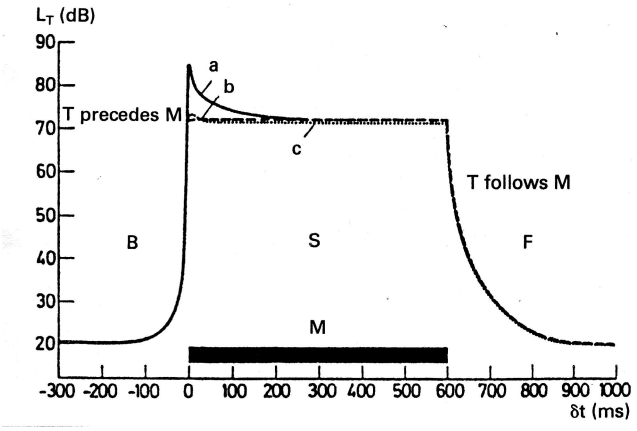

Figura 2.9: Enmascaramiento anterior y posterior [13]

- El enmascaramiento permanece unos 120 ms una vez se suprime M.

- **Enmascaramiento por ruido de ancho de banda estrecho: Banda Critica y Resolucion en Frecuencia del Sistema Auditivo**
  Cuando un tono es enmascarado por ruido blanco, sólo las frecuencias ruidosas cercanas a la del tono enmascarado juegan un importante papel.
  Los datos muestran que el mínimo ancho de banda δf que permite enmascarar un tono a su frecuencia central f guarda un relación fisiológica con las discriminaciones en frecuencia que se realizan en la cóclea. Este ancho de banda mínimo se designa banda crítica[13].
  El sistema auditivo parece funcionar como un banco de filtros paso banda cada uno de ellos teniendo un cierto ancho de banda δf_c. Los anchos de banda críticos varían según la frecuencia[13], el resultado de experimentos que avalan esta aseveración puede apreciarse en la figura 2.10.
  Ciertos investigadores apuestan por la importancia de los intervalos críticos, y utilizan en este contexto una unidad mayor que el mel pero de las mismas dimensiones, el Bark (1 Bark=100 mels). El interés práctico del Bark radica en que representa a lo largo de la cóclea las separaciones frecuenciales correspondientes a cada banda crítica[13].

- **Efectos Temporales**

  - **Enmascaramiento Posterior**
    Mediante experimentos ([13]) se han obtenido los siguientes resultados:

    1. Para el caso $f_M = f_T$: Se produce un elevamiento del umbral

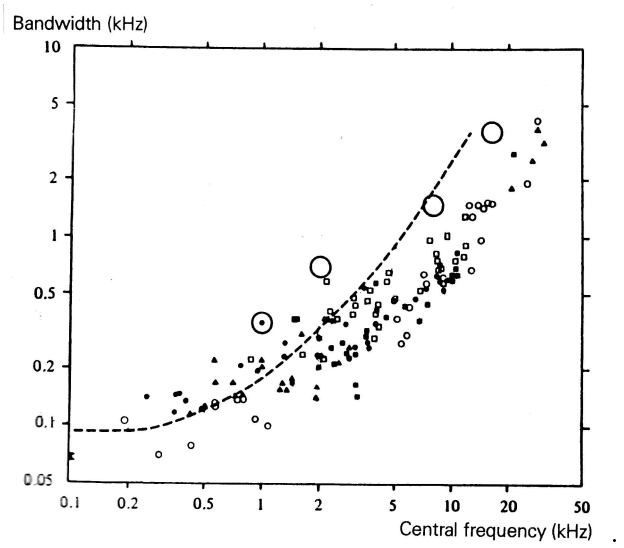

Figura 2.10: Variación del ancho de banda crítica en función de la frecuencia central [13]

que sugiere la existencia de una fatiga auditiva, o alternativamente, demuestra una pérdida en la sensibilidad del sistema auditivo debida a la exposición previa a un sonido enmascarador

2. En el caso $f_M \neq f_T$: La curva $L_T = \Phi(L_M)$ depende de la separación entre frecuencias, pero la forma es esencialmente la misma.

- **Fatiga y Adaptación Auditiva**
  Definimos la fatiga auditiva como una modificación de la sensibilidad auditiva tras una exposición prolongada a un sonido[13]. Se produce una pérdida que esta caracterizada por:
  1. Un ascenso del umbral.
  2. Una vuelta a la sensibilidad normal mucho más lenta que en el caso del enmascaramiento posterior
- **Adaptación a Sonidos Modulados**
  Existe una adaptación a los sonidos AM y/o FM. La exposición a un sonido con una cierta modulación durante, por ejemplo, 10 segundos genera un deterioro en el umbral de detección de la modulación durante unos segundos. Esto es debido a que el sistema auditivo cuenta con un cierto numero de "canales" dedicados a la deteccion de modulaciones en los sonidos. Estos "canales" juegan un importante papel en la detección de cambios temporales[13].
- **Influencia de la Duración de los Estímulos en el Umbral: Sonidos Breves**

Los efectos en los umbrales que han sido visto hasta el momento se han considerado independientes a la duración del estímulo. Esto es cierto a partir de una cierta duración.
En audición, el umbral absoluto de intensidad detectable I de un estímulo se considera constante para cualquier duración del estímulo siempre que ésta supere cierto mínimo, por debajo de éste disminuye progresivamente según la duración d se hace menor[13]. Para estos valores pequeños (menores a 180 ms) se cumple la relación:

$$L \cdot d = C \tag{2.4}$$

Siendo C una constante. También dependen de esta duración los umbrales diferenciales mínimos δL y δf.

- **Influencia de la Duración del Estímulo en el Umbral Máximo**
  Para señales de duración entre 100 y 200 ms la sensación subjetiva de volumen sonoro disminuye al disminuir la duración.
  De igual modo, el tono percibido ha sido estudiado en estas circunstacias. Observandose tres casos:
  - Duraciones muy cortas: sólo se percibe un "click", sin ninguna cualidad tonal
  - Duraciones medias: se percibe un "click", pero con cierta tonalidad
  - Duraciones más largas: se percibe un "click" inicial, seguido de una sensación tonal que termina en otro "click".

  Estas sensaciones son también dependientes de la frecuencia. Para frecuencias por debajo de 1 kHz es necesario un mínimo número de ciclos de onda para percibir el sonido, pero para frecuencias mayores de 1 kHz es necesario una duración mínima del sonido[13].

Efectos Biaurales

- **Suma Biaural y Umbral Absoluto**

  - El umbral biaural de detección de un sonido es, en general, menor que en el caso monoaural.
  - En principio, si ambos oídos tuviesen el mismo umbral monoaural y se produjese a cierto nivel del sistema nervioso central una suma perfecta de los efectos de los dos oídos, el umbral biaural seria el 50 % del monoaural (3dB menos).
  - En la mayoría de los casos el umbral monoaural de los dos oídos es distinto, a veces diferencias mayores a 6dB[13].

- **Umbral Diferencial de Intensidad**
  El umbral diferencial de intensidad en el caso biaural es menor que en el caso monoaural[13].

- **Suma de Intensidades Biaurales**
  Un sonido aplicado en ambos oídos parece más alto que cuando es aplicado a uno solo. La forma en que se suman las intensidades dependerá de las frecuencias de los sonidos aplicados a cada oído, siendo máxima en el caso $\delta f = |f' - f| = 0$[13].

- **Suma Interaural e Inhibición durante el Enmascaramiento**
  Cuando un sujeto es expuesto a una señal ruidosa enmascaradora débil, el umbral biaural es menor que el monoaural. Produciéndose el efecto contrario si la señal enmascaradora es fuerte, este efecto, denominado Inhibición Interaural es particularmente severo a bajas frecuencias[13].

- **"Diplacusis": Diferencias Interaurales en Tono Percibido**
  Se define la diplacusis [13] como la percepción de distintos tonos cuando sonidos de las misma frecuencia son aplicados a uno u otro oído.

Sensaciones Tonales

Se tratarán a continuación ciertos estudios sobre la naturaleza física del sonido y las propiedades fisiológicas de los receptores auditivos.

- **Percepción de Sonidos Reales**

  - **Atributos de Tonos Puros**
    El concepto de tonalidad y consonancia de octavas es aceptado como un atributo del sonido, aunque no es lo mismo que tono percibido. La distinción de nivel del tono 'alto-bajo' representa una variación monótona. Mientras que la idea de tonalidad y consonancia en octavas conforma una relación cíclica entre frecuencias[13]. Además de volumen sonoro y tono percibido los tonos puros presentan otra serie de atributos subjetivos difíciles de cuantificar. Destacamos los tres siguientes: claridad, volumen, densidad.
    Otra importante cualidad de los tonos puros es rugosidad, que se experimenta cuando un tono puro es modulado en amplitud a una frecuencia mucho menor que la suya. De los experimentos sobre este fenómeno resaltamos tres resultados[13]:
    - La impresión subjetiva de rugosidad(r) es proporcional al cuadrado del índice de modulación $m(m = \delta p/p)$:

$$r = rm^2 \tag{2.5}$$

- La intensidad del sonido modulado no parece jugar un papel importante.
- La rugosidad varía con la frecuencia de modulación.

- **Escuchar Dos Tonos Puros Simultaneamente**
  Debemos examinar también el conjunto de percepciones subjetivas que se producen cuando dos tonos puros sinusoidales son escuchados simultaneamente, siendo sus intensidades parecidas y cuyas frecuencias f_1 y f_2 guardan alguna relación.
  Cuando la diferencia entre las frecuencias es de unos 2 ó 3 Hz, se oye un único tono cuya intensidad varía periodicamente. Este sonido es percibido como un batido a una frecuencia $\delta f = f_1 - f_2$, este batido cesa para valores $\delta f \geq 6Hz$. Para diferencias de frecuencias mucho mayores los tonos parecen oirse de forma más consonante o disonante en función del ratio de frecuencias. Siendo muy consonante para los ratios 2, 5/4, 4/3, 3/2[13].

- **Tonos Puros y No-linealidades**
  Existe otro fenómeno que sólo afecta cuando la intensidad del tono supera cierta intensidad. A niveles altos de intensidad sonora se producen distorsiones que son universalmente atribuidas a no-linealidades del sistema auditivo.
  Cuando un sujeto es expuesto a un único tono puro, libre de distorsiones armónicas, por encima de cierta intensidad oye armónicos. A este fenómeno se le conoce como "overtones" o armónicos aurales.
  Cuando escuchamos un tono puro, de frecuenmcia f_1, y otro tono de frecuencia f_2 es añadido con la misma intensidad, la percepción de volumen del tono primero se hace menor cuanto mayor se hace la del segundo.
  Bajo otras condiciones, no solo los dos primeros tonos son oídos, sino también tono combinados con frecuencias $mf_1 \pm nf_2$. Los tonos de frecuencias $mf_1 - nf_2$ son llamados "difference tones" y los de $mf_1 + nf_2$ "summation tones". En general los "difference tones" son mucho más evidente que los otros. Siendo dos tonos combinados especialmente notables: el "simple difference tone" $f_2 - f_1 (f_2 > f_1)$, y el "cubic difference tone" $2f_1 - f_2$.
  Estos efectos no-lineales son producidos por el sistema auditivo, si bien no se conoce bien en qué parte del mismo. Aunque se les relaciona con la selectividad en frecuencia[13].

- **Percepción de Sonidos Complejos**
  Los sonidos complejos pueden estar formados por una mezcla de tonos puros, o por un tono fundamental y sus armónicos[13].
  Georg Simon Ohm estableció una primera relación entre los componentes de un sonido complejo que son subjetivamente percibidos

y los componentes armónicos que se observan mediante un análisis de Fourier[13].
Esta y posteriores ideas dieron lugar a la teoría de que el sistema auditivo puede verse, al menos en sus primeras etapas, como un banco de filtros paso-banda en paralelo. Existiendo una mayor resolución en frecuencia para los armónicos a bajas frecuencias, y una menor resolución para los de frecuencias altas.
Uno de los atributos de los sonidos complejos es su timbre, la impresión subjetiva que nos permite distinguir entre dos sonidos con el mismos volumen y tono percibido. El timbre es un atributo multidimensional que depende de distintos factores. Uno de estos factores es el numero total de armónicos, y sus amplitudes. También es de gran importancia la fase de los componentes. Es claro que la fase de los distintos componentes afecta a la forma de onda resultante[13].

2.3. Modelos del Sistema Auditivo

En esta sección vamos a adelantar los tres modelos auditivos que han servido de partida para los distintos experimentos de los que consta el Proyecto.

Los conocimientos expuestos en las secciones anteriores nos permitirán entender más fácilmente los diferentes modelos.

Los tres modelos explicados en esta sección (Lyon, ERB y Seneff) modelan fundamentalmente la cóclea, aunque alguno también modele otras partes del oído. Por tanto tomarán como entrada la señal auditiva, o una representación de la misma, y darán como salida un cocleograma que es la probabilidad de activación de las neuronas del nervio auditivo a lo largo del tiempo[9].

2.3.1. Modelo de Lyon

Richard F. Lyon desarrolló un modelo coclear basado en los conocimientos sobre el funcionamiento de la cóclea[9]. Este modelo describe la propagación del sonido en el oído interno y la conversión de la energía acústica en representaciones neuronales[23].

Este modelo no intenta describir literalmente cada estructura de la cóclea, sino que considera toda ella una caja negra. El sonido entra en la cóclea vía la ventana redonda y es convertido en descargas neuronales que viajan a través del nervio auditivo[23].

El modelo coclear descrito por Lyon combina[23]:

- Una serie de filtros que modelan la onda de presión viajera.

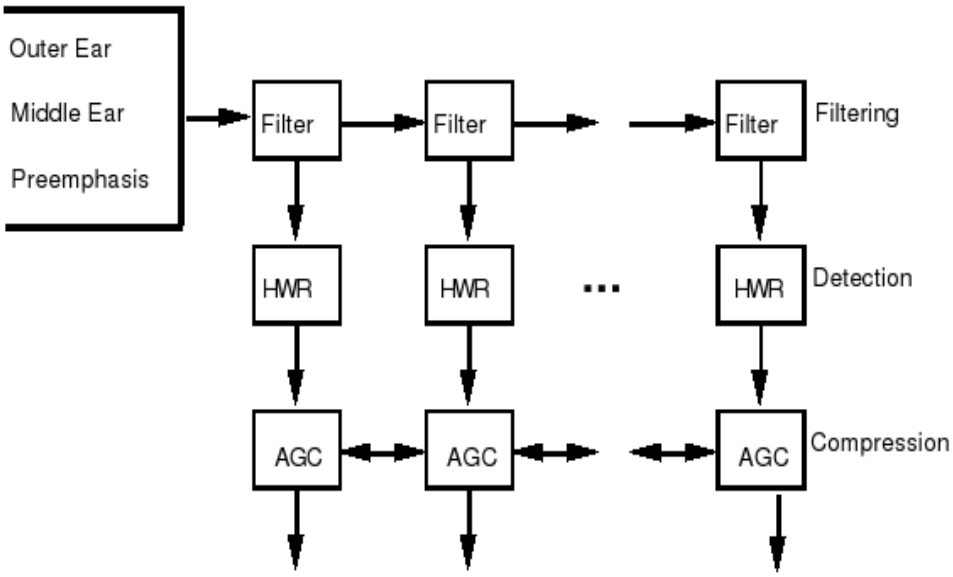

Figura 2.11: Esquema básico del modelo de Lyon [23]

- Rectificadores de Media Onda, en inglés Half Wave Rectifiers (HWR), para detectar la energía de la señal. Actúan como lo harían las Células Ciliadas Internas[9].
- Distintas etapas de Control Automático de Ganancia, en inglés Automatic Gain Control (AGC). En el oído esta acción la realizan las Células Ciliadas Externas.

Un esquema de este modelo puede observarse en la figura 2.11.

En el modelo de Lyon existe una etapa de preénfasis inicial seguida por una cascada de etapas de filtros auditivos. Ese filtro de preénfasis es utilizado para modelar los efectos del oído externo y medio, para conseguirlo se utiliza un filtro paso alto con una frecuencia de corte de 300 Hz. Esta seguido por un diferenciador y un compensador de alta frecuencia comunes a todas las etapas[23].

En cada punto de la cóclea la onda acústica es filtrada por un filtro "notch". Cada filtro "notch" opera en una frecuencia satisfactoriamente baja, de forma que el efecto global es un filtrado paso bajo gradual. Un resonador adicional (filtro paso banda) deja pasar una pequeña parte de la energía de la onda viajera y modela la conversión del movimiento de la membrana basilar que es detectado por las Células Ciliadas Internas[23].

La señal de cada etapa de filtrado es una representación paso banda de la señal de audio original. Esta representación transcurre por un rectificador de media onda y una etapa de control de ganancia.

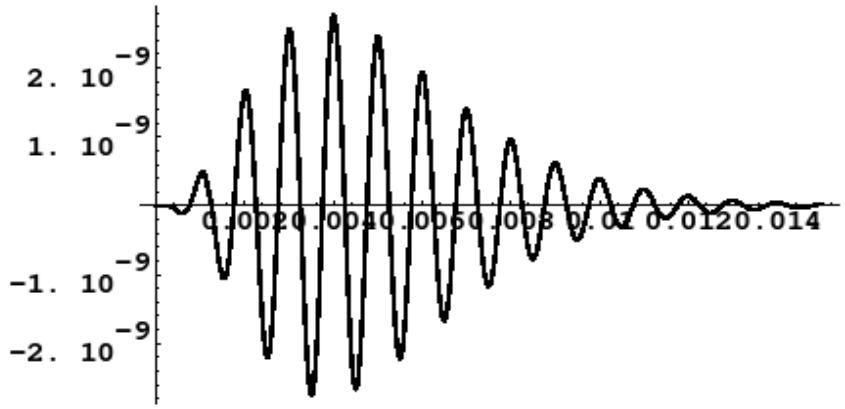

Figura 2.12: Gammatone [21]

2.3.2. Modelo ERB

El modelo ERB está basado en los trabajos de Patterson y Holdworth sobre la cóclea[21].

Se basa en una matriz de filtros paso banda independientes. Cada uno de ellos sintonizados a una frecuencia diferente. En el modelo de Patterson el ancho de banda de cada filtro coclear esta descrito por un Ancho de banda Rectangular Equivalente, en inglés "Equivalent Rectangular Bandwith" (ERB)[21].

El ERB es una medida psicoacúsica del ancho del filtro auditivo para cada punto a lo largo de la cóclea. que hemos definido anteriormente como Bandas Críticas, usaremos ambos términos indistintamente. Un filtro de bandas críticas o ERB modela la señal que está presente en una única célula del nervio auditivo o un canal [21].

El modelo de Patterson esta basado en filtros "gammatone" con respuestas al impulso como el de la figura 2.12.

En el modelo de Patterson no se especifica la separación frecuencial entre canales[21]. Ni se realiza control de ganancia ni adaptación.

2.3.3. Modelo de Seneff

El modelo de Seneff es similar al de Lyon y trata de capturar las características fundamentales extraidas por la cóclea.

Consta de tres bloques, de los cuales sólo explicaremos los dos primeros ya que el tercero no es relevante para el Proyecto.

La señal de voz es prefiltrada para eliminar frecuencias muy altas y muy bajas.[6]

1. Primer bloque: un banco de filtros lineales de bandas críticas de 40 canales. Estos canales fueron diseñados para coincidir con los datos fisiológicos disponibles.[6].

2. Segundo bloque: No lineal, captura las características predominantes de la vibración de la membrana basilar. La salida de esta etapa representa la probabilidad de disparo de los haces nerviosos en función del tiempo.[6].

2.4. Conclusiones

A lo largo del capítulo hemos introducido de forma detallada el funcionamiento del oído y el aparato fonador. Hemos hecho especial incapié en la percepción subjetiva que realizamos de los sonidos en función de su intensidad y frecuencia. También se ha presentado la morfología del sistema auditivo y se ha explicado qué procesado realiza cada una de las partes del sistema auditivo. Siendo la cóclea y el nervio auditivo los lugares en los que se producen los principales procesos como son el análisis espectral, el enmascaramiento o la resolución en frecuencia.

Como se comentó en la sección 1.1 diseñar Sistemas de Reconocimiento de Habla que tengan en cuenta el funcionamiento real del oído humano es de gran utilidad y supone una importante ventaja.

Saber de qué manera el oído procesa la onda sonora hasta convertirla en impulsos nerviosos o la adaptación y sintonización de los sonidos que se produce en el nervio auditivo ha permitido desarrollar complejos Modelos Auditivos como los de la seccion 2.3.

Estos modelos auditivos pueden suponer una mejora frente a los modelos clásicos que conforman la base de las parametrizaciones plp o mfcc (ver la sección 1.1.2). Comprobar si parametrizaciones basadas en modelos como el de Lyon o ERB suponen una ventaja frente a las parametrizaciones clásicas es uno de los objetivos fundamentales de este proyecto y se tratará de forma experimental en el capítulo 5.

Capítulo 3

Procesado Morfológico

3.1. Morfología matemática

La morfología matemática, o matemática morfológica, forma parte de la teoría de conjuntos. Se utilizó originariamente para el procesado de imágenes. Y, aunque en la actualidad se ha extendido su uso a otras áreas, ésta sigue siendo la principal. No obstante, este Proyecto hace uso de ella como una herramienta más en el procesado de la señal de voz[1].
Sus dos procesos fundamentales son las operaciones no lineales de erosión y dilatación[12]:

$$\text{Erosion } (f \ominus b)(x) = \text{ínf}\{f(y) - b(y-x) | y \in \Re^n\} \tag{3.1}$$

$$\text{Dilatacion } (f \oplus b)(x) = \sup\{f(y) + b(x-y) | y \in \Re^n\} \tag{3.2}$$

Podemos generalizar la dilatación a dilatación tangencial:

$$\text{Dilatacion tangencial } (f \breve{\oplus} b)(x) = \underset{y}{sup}\{f(y) + b(x-y)\} \tag{3.3}$$

$$\underset{y}{sup}\, f(y) = \{f(z) | \nabla f(z) = 0\} \tag{3.4}$$

La morfología matemática, al igual que la Teoría de Sistemas Lineales, es una herramienta para describir ciertos aspectos de la naturaleza. Provee una vía alternativa de combinar señales distinta a la habitual (lineal)[12].

[1] Para ver cómo y para qué ha sido utilizado el procesado morfológico en este proyecto diríganse al capítulo 5

	Teoría Lineal	**Teoría Morfológica**
Transformada Canónica	$F[f](u)$	$S[f](u)$
Teorema de la 'Convolución'	$F[f * b] = F[f]F[b]$	$S[f \breve{\oplus} b] = S[f] + S[b]$
Kernel Canónico	$\frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{<x,x>}{2\sigma^2}}$	$-\frac{<x,x>}{4t}$

Tabla 3.1: Comparación TMS vs TLS[12].

3.2. Transformada Slope o de la Pendiente

La Transformada de Fourier(TF) juega un papel importantísimo en el procesado lineal de señales. Recordemos su fórmula:

$$F[f](u) = \int_{\Re^n} f(x) e^{-j2\pi <u,x>} dx \tag{3.5}$$

La Transformada de Fourier nos provee de una descripción en el dominio de la frecuencia. Además, en este dominio la convolución temporal de señales se comporta como una multiplicación. Lo que facilita enormemente tareas como el filtrado.

$$F[f * g] = F[f]F[g] \tag{3.6}$$

El equivalente a la TF en el procesado morfológico es la Transformada Slope (TS), o transformada de la pendiente[12]. Cuya fórmula es:

$$S[f](u) = \underset{x}{stat}\{f(x) - < u, x >\} \tag{3.7}$$

y su inversa:

$$S^{-1}[g](x) = \underset{u}{stat}\{g(u) + < u, x >\} \tag{3.8}$$

Bajo el dominio de la pendiente la dilatación se convierte en suma:

$$S[g \breve{\oplus} f] = S[g] + S[f] \tag{3.9}$$

Los paraboloides son las funciones estructurales en Teoría Morfológica de la Señal (TMS), como las Gaussianas lo eran en Teoría Lineal de la Señal (TLS)[12].

$$b(x,t) = -\frac{< x, x >}{4t} \tag{3.10}$$

siendo $(t > 0)$.

La tabla 3.1 sugiere que existe una conexión entre TLS y TMS. De hecho morfología matemática es en esencia Teoría de Sistemas Lineales, pero haciendo uso de un algebra específico[12].

	Conjunto	**Adición**	**Multiplicación**
Algebra 'plus-prod' $\Re$	$\Re$	$+$	$\times$
Algebra 'max-plus' $\Re_{max}$	$\Re \cup \{-\infty\}$	máx	$+$
Algebra 'min-plus' $\Re_{min}$	$\Re \cup \{+\infty\}$	mín	$+$

Tabla 3.2: Definición de los algebras 'max-plus' y 'min-plus'[12]

3.3. Algebra plus-prod vs Algebra max-plus

Mientras TLS hace uso del álgebra estandard, denominado 'plus-prod', TMS esta basado en el álgebra 'max-plus' o 'min-plus'.
Estas formas de álgebra alternativas se forman a partir del álgebra 'plus-prod' añadiendo $+\infty$ ó $-\infty$ al eje real y sustituyendo la operación de suma por la de máximo o mínimo, y la multiplicación por la de suma[12].
Estos cambios se reflejan más claramente en la tabla 3.2.
Veamos a continuación cómo queda la convolución del álgebra lineal bajo estos nuevos álgebras[14].
En álgebra lineal teníamos:

$$(f * g)(x) = \int_{\Re^n} f(x-y)g(y)dy \tag{3.11}$$

En $\Re_{max}$:

$$(f \underset{d}{*} g)(x) = \sup_{y \in \Re^n} \{f(x-y) + g(y)\} = \sup_{y \in \Re^n} \{f(y) + g(x-y)\} \tag{3.12}$$

En $\Re_{min}$:

$$(f \underset{e}{*} g)(x) = \inf_{y \in \Re^n} \{f(x-y) + g(y)\} = \inf_{y \in \Re^n} \{f(y) + g(x-y)\} \tag{3.13}$$

Que coinciden con las operaciones antes descritas de erosión y dilatación.

$$(f \oplus g)(x) = \sup_{y \in \Re^n} \{f(y) + g(x-y)\} = (f \underset{d}{*} g)(x) \tag{3.14}$$

$$(f \ominus g)(x) = \inf_{y \in \Re^n} \{f(y) - g(y-x)\} = (f \underset{e}{*} \bar{g})(x) \tag{3.15}$$

Siendo: $\bar{g}(x) = -g(-x)$.
Por tanto, observamos cómo las operaciones no lineales de erosión y dilatación pueden identificarse con convoluciones lineales en los álgebras 'max-plus' y 'min-plus'.

3.4. Transformada de Legendre-Fenchel

La Transformada Slope genera funciones vectoriales, pero existe una variante que nos permite obtener y trabajar con funciones escalares. Esta es la Transformada de Legendre-Fenchel.
La Transformada de Legendre-Fenchel también es conocida como operación de conjugación, o conjugada convexa[3]. Esta definida por la siguiente fórmula:

$$f^*(x) = \sup_{t \in \Re^n} \{< t, x > -f(t)\} \tag{3.16}$$

Por otra parte, en TLS definimos la Transformada de Laplace multivariante como[12]:

$$L[f](x) = \int_{\Re^n} e^{<x,y>} f(y) dy \tag{3.17}$$

Y como ocurría con la TF la convolución se convierte en multiplicación.

$$L[f * g](x) = L[f]L[g] \tag{3.18}$$

Existe una importantísima relación entre estas dos transformadas. La conjugación, o Transformada de Legendre-Fenchel, de f interpretada en el contexto del álgebra 'max-plus' corresponde al logarítmo de la Transformada de Laplace multivariante de e^{-f} en el álgebra estandar[12]. Esta relación se demuestra a continuación:

$$f^*(x) = \sup_{y \in \Re^n} (< y, x > -f(y)) = \lg \sup_{y \in \Re^n} (e^{<y,x> - f(y)}) \tag{3.19}$$

Cuya representación en álgebra 'plus-prod' queda:

$$\lg \int_{\Re^n} e^{<y,x> - f(y)} dy = \lg \int_{\Re^n} e^{<y,x>} e^{-f(y)} dy = \lg L[e^{-f}](x) \tag{3.20}$$

3.5. Transformada de Cramer

Por último, definiremos una última transformada. La Transformada de Cramer (TC), que es la combinación del logaritmo de la Transformada de Laplace con su equivalente morfológico, la operación de conjugación o Transformada Legendre-Fenchel[12][3][4].
La TC esta definida por la funcion:

$$C[f] = (\lg L[f])^* \tag{3.21}$$

3.5.1. Comparación con Cepstrum

Si observamos la ecuaciones 3.21 y 1.3 veremos importantes similitudes entre ambas.

$$C[f] = (\lg L[f])^* \tag{3.22}$$

$$Cepstrum[f] = TF^{-1}(\lg |X(\omega)|) \tag{3.23}$$

Como acabamos de definir, la Transformada de Cramer se calcula haciendo la Transformada de Legendre-Fenchel sobre el logaritmo de la Transformada de Laplace de la función en cuestión. Mientras que el Cepstrum lo calculabamos haciendo la Transformada Inversa de Fourier (que es una operación prácticamente idéntica a la Transformada de Fourier) sobre el logaritmo del módulo de la Transformada de Fourier de la función en cuestión.

Estas similitudes nos permitirán calcular "algo parecido" al Cepstrum de una señal simplemente realizando su Transformada de Cramer. Esta importante característica se encuentra en la base de algunos de los experimentos. Profundizaremos sobre ello más adelante, en la sección 5.3.

Capítulo 4

Entorno de Experimentación

Como se explicó en el capítulo el proyecto tratará de analizar cuales de entre una serie de parametrizaciones son mejores en según qué circunstancias. Para ello es necesario contar con un entorno de experimentación estable.

Este entorno de experimentación consta de un "corpus" o base de datos de voz, y un Sistema de Reconocimiento de Habla. El "corpus" nos provee la materia prima para las distintas parametrizaciones y el Sistema de Reconocimiento de Habla un Sistema de Pruebas para poder evaluar estas parametrizaciones.

A lo largo del capítulo se dan las especificaciones técnicas y de utilización de estos dos elementos.

4.1. La Base de Datos

Para la realización de este proyecto hemos utilizado como base de datos, o corpus, "ISOLET v1.3".

Esta ha sido la base de datos utilizada porque a pesar de su reducido y manejable tamaño incluye todas las letras del alfabeto inglés[7].

ISOLET es una base de datos que contiene 7800 letras del alfabeto inglés pronunciadas de forma aislada. Cada letra ha sido pronunciada y grabada 2 veces por cada uno de los 150 hablantes que intervinieron. Lo cual hacen un total de, aproximádamente, 85 minutos de habla.

Las grabaciones se realizaron en laboratorio con un micrófono cancelador de ruidos. La frecuencia de muestreo es de 16000 Hz.

Los hablantes que participaron en las grabaciones tenían el inglés como lengua materna y una edad comprendida entre los 14 y los 72 años. La mitad eran hombres y la otra mitad mujeres[1].

Esta base de datos limpia fue contaminada con diferentes tipos de ruido para poder realizar los experimentos en condiciones menos ideales que las del laboratorio. Los ruidos fueron obtenidos de la colección "RSG-10" [24]. Para llevar a cabo esta contaminación se siguieron tres principios fundamentales[7]:

- Distintos tipos de ruido.
- Distintas Relaciones Señal a Ruido (SNR).
- Condiciones de entrenamiento limpia y ruidosa.

La base ISOLET fue dividida en 5 partes iguales, contaminándose todas ellas con tres ruidos diferentes ("Speech babble", "Factory floor noise 2", "Car interior noise"). Y a continuación cada una de las 5 partes con 1 de los siguientes ruidos ("Pink noise", "F-16 cockpit noise", "Destroyer operations room noise", "Military vehicle noise", "Factory floor noise 1")[7].

Por cada una de estas partes existe un directorio con los diferentes archivos que codifican la voz.

4.2. El Sistema de Pruebas

El sistema de pruebas que se ha utilizado en este Proyecto ha sido el "ISOLET Testbed", desarrollado por el ICSI.

Este sistema de pruebas implementa una aproximación al modelo híbrido de reconocimiento de voz de [11], que ya se explicó en la sección 1.2.2, por tanto no nos extenderemos aquí con las explicaciones teóricas. La razón por la que se elige un sistema híbrido basado en MLP en lugar de en Gausianas es que los MLP han dado algunas veces mejores resultados con parametrizaciones nuevas[15].

Las distintas herramientas necesarias han sido obtenidas de "SPRACHcore-nogui-2004-08-26", de "quicknet3" y del conjunto de librerías "dpwelib-2006-04-19". Todas ellas pueden obtenerse de manera gratuita a través de la página web del ICSI.

Como explicamos en la sección anterior la base de datos de habla esta dividida en 5 secciones, o "fold" en inglés, limpias y otras 5 ruidosas. El orden en que los ficheros de las distintas secciones debe ser leido es siempre el mismo y se define en los archivos "noisy.wav.files.rand" y "clean.wav.files.rand".

El primer paso de todos los experimentos es parametrizar esta base de datos. Para ello utilizaremos los modelos clásicos que nos provee SPRACHcore ("plp" y "mfcc") o las distintas parametrizaciones desarrollados en este Proyecto[1]. Con ello obtendremos 2 ficheros (uno limpio y otro ruidoso) de

[1]Las distintas parametrizaciones se detallan en el capítulo 5

tipo "pfile". El nombre que reciben estos archivos debe seguir el siguiente patrón: estado("clean" o "noisy")-parametrización.pfile. Por ejemplo: "clean-plpD2.pfile" o "noisy-mrmpecv5.6.pfile".

Estos archivos "pfile" son procesados mediante la herramienta "feacat" para ajustar el numero de tramas por expresión a la cantidad de referencia que se especifica en el fichero "all.deslen".

Los ficheros "pfile" deben normalizarse para su correcta utilización. La herramienta "norms" nos provee esta funcionalidad.

Una vez tenemos la base de datos parametrizada y normalizada puede comenzar el entrenamiento del reconocedor.

Nuestro Sistema Automático de Reconocimiento de Habla esta formado por 1600 nodos ocultos. Este valor es ajustable, sin embargo lo hemos mantenido fijo para poder comparar nuestros resultados con los experimentos de referencia con los que hemos trabajado, que contaban con 1600 nodos.

Utilizaremos un contexto entre 5 y 9, en función del experimento.

El número de parámetros de los vectores de características variará en las distintas parametrizaciones.

Utilizaremos las cinco secciones para entrenamiento. La herramienta "qnstrn", ejecutada mediante el script "train", es la encargada de entrenar el sistema. Para entrenar se llevan a cabo 5 iteraciones. En cada una de ellas son utilizadas 4 secciones para entrenar y la restante como referencia de parada. Esta referencia va cambiando en las siguientes iteraciones

Después de entrenado, tanto para limpio como para ruidoso, con el script "tune" se fijan los mejores parámetros de decodificación. Utilizamos para ello únicamente el "fold1".

A continuación se procede a "testear" el sistema. Para ello SPRACHcore provee la herramienta "noway" a la cual nosotros accedemos con el script "testPP".

Por último sólo nos queda calcular la tasa de error. Para este cálculo sólo tenemos en cuenta las 4 últimas secciones, no el "fold1". Aunque lo podríamos hacer a mano a partir de los datos obtenidos en la fase de test utilizamos el script "signifForFoldsPP" ya que de esta forma se nos presenta la información de una forma más cómoda. Además nos da otros datos que nos servirán para calcular la significancia estadística de los resultados obtenidos. Estos últimos pasos quedan más explicados en el capítulo 5.

Todo este sistema de pruebas, así como la base de datos ha sido montado en un PC portatil. Sobre un sistema operativo Ubuntu 6.10.

Capítulo 5

Experimentos y Resultados

Ya vimos en la introducción que el objetivo del proyecto es comparar las parametrizaciones clásicas de habla, "plp" y "mfcc", con un conjunto de nuevos modelos.

Esta comparación se realizará a tres niveles distintos, en función del tipo de datos que se hayan usado en el entrenamiento y la fase de test[1]. Pudiendo ser:

- **Ajustado:**
  - Entrenamiento limpio y test limpio.
  - Entrenamiento ruidoso y test ruidoso.
- **Desajustado:**
  - Entrenamiento limpio y test ruidoso.

Tomaremos en todos los casos como referencia los datos que se obtengan para la parametrización "plp", ver sección 5.1, ya que es la parametrización clásica con mejores resultados. Y compararemos con estos datos calculando el porcentaje de mejora obtenido respecto a "plp". Además del porcentaje de mejora, se aportará en cada caso el grado de significancia estadística del dato en cuestión.

El grado de significancia estadística nos da una valiosa información, en función del numero de errores y de los parámetros de la base de datos, sobre la probabilidad de haber obtenido ese resultado de forma casual. Esto es, nos cuantifica la confianza en nuestros resultados experimentales. El cálculo de estos datos se ha realizado mediante el test de significancia estadística de McNemar, a través de [5] donde además se puede ampliar la información sobre este test.

[1] Más información sobre el Sistema de Pruebas en la sección 4.2.

Los experimentos realizados se han dividido en tres grandes bloques:

1. Experimentos de Referencia, en la sección 5.1.
2. Nuevos Modelos Auditivos, en la sección 5.2.
3. Transformada de Cramer, en la sección 5.3.

En primer lugar se han obtenido, mediante las herramientas del SPRACHcore de ICSI, unos valores de referencia para las parametrizaciones clásicas "plp" y "mfcc".

A continuación se han desarrollado nuevas parametrizaciones, a partir de distintos modelos auditivos. Estas parametrizaciones modelan el oído de manera distinta a la de "plp" o "mfcc", si bien en el último paso se ha utilizado un análisis similar al "cepstral" para obtener los vectores de parámetros.

El último grupo de experimentos ha tratado de dar un paso más. Al igual que los casos anteriores hace uso de los mismos nuevos modelos auditivos pero en lugar de hacer un análisis "cepstral" mediante la Transformada de Fourier se ha utilizado la Transformada de Cramer directamente para calcular los vectores de características. Este cambio se introducía de forma teórica en la sección 3.5.1 y se amplía de forma práctica en la 5.3.

Los modelos auditivos que se han probado en este Proyecto han sido tres:

- Modelo de Lyon
- Modelo ERB
- Modelo de Seneff

Sobre sus principios teóricos ya hablamos en la sección 2.3. Las implementaciones concretas utilizadas en este proyecto han sido las desarrolladas para MATLAB que pueden encontrarse en el "AuditoryToolbox" de Malcolm Slaney[22]. Además en el anexo B hay más información sobre "AuditoryToolbox".

Para llevar a cabo las parametrizaciones de los nuevos modelos ha sido necesario generar y ejecutar código MATLAB. La versión de MATLAB utilizada ha sido la 7.1.0.183 (R14) Service Pack 3, con la toolbox mencionada antes: "AuditoryToolbox".

5.1. Experimento de Referencia

Para comprobar si todas las herramientas del Sistema de Pruebas, así como la Base de Datos, funcionaban correctamente se llevaron a cabo dos experimentos de prueba.

	Clean - Clean	**Clean - Noisy**	**Noisy - Noisy**
PLP	3,6	53,7	17,4
MFCC	4,1	58,9	18,6

Tabla 5.1: Tasa de Error, por palabra. Experimentos de Referencia.

	Clean - Clean	**Clean - Noisy**	**Noisy - Noisy**
% Mejora	-13,88	-9,68	-6,89
% Prob. Hipótesis Nula	0,76	despreciable	despreciable

Tabla 5.2: Porcentaje de Mejora y Significación Estadística para MFCC.

Estos experimentos habían sido realizados previamente por los desarrolladores del SPRACHcore, por lo que conocíamos los resultados previamente.

Además, los resultados de estos experimentos nos servirían como referencia para comparar los resultados de los experimentos con las nuevas parametrizaciones.

Estos dos experimentos se realizaron con un contexto de 5 y, al igual que todos los demás, con 1600 nodos en la capa oculta. Los vectores de parámetros estaban constituidos por 39 elementos.

Para cada tramo enventanado se calcularon 12 parámetros más el valor de la energía. De estos 13 parámetros se calcularon las derivadas y las derivadas segundas, obteniéndose los 39 parámetros finales.

Se utilizó un tamaño de ventana de 25ms.

Las parametrizaciones que se realizaron en estos experimentos fueron:

- plp(sin RASTA)
- mfcc

Los resultados obtenidos fueron los reflejados en la tabla 5.1.

Los resultados de la tabla 5.1 no coinciden exactamente con los obtenidos por los desarrolladores para mfcc. Tras una serie de comprobaciones se llegó a la conclusión de que se debía a diferencias de procesador. En cualquier caso, para los resultados y conclusiones del Proyecto se tendrán siempre en cuenta los resultados puntuales obtenidos durante el mismo.

Como ya hemos comentado utilizaremo los datos de plp como referencia para el resto de parametrizaciones. Incluimos por tanto en la tabla 5.2 los datos de mejora porcentual y del test significación estadística para mfcc.

5.2. Nuevos Modelos Auditivos

Las parametrizaciones utilizadas en el Proyecto hacen uso del "Auditory-Toolbox" [22].

Hemos utilzado tres modelos auditivos diferentes y se han variado los parámetros de estos modelos, pero la naturaleza de los experimentos ha sido siempre la misma.

La señal de entrada completa, proviniente de la base de datos , es filtrada con un filtro paso alto. Pasa por el modelo auditivo en cuestión. Como salida de este procesamiento se obtienen un vector de versiones filtradas, según los distintos modelos, de la señal de origen.

Se calcula la Transformada del Coseno de este vector y nos quedamos con los 12 primeros coeficientes de esta transformada. Mediante este procedimiento hacemos un procesado similar al "cepstrum", nos quedamos con la parte de la señal que encierra la información fonética de la señal.

A estos 12 parámetros le añadimos otro que es la energía de la señal. Por último, calculamos la derivada y la derivada segunda de los 13 parámetros, obteniendo finalmente 39 parámetros por vector de características.

Para todos los modelos se ha experimentado con contexto 5 y 7.

A continuación se explican más en profundidad cada uno de los modelos, y las distintas variantes, con los que se ha experimentado.

Para poder comprender mejor las diferencias entre los distintos modelos auditivos de esta y la siguiente sección se dan los valores de los parámetros con los que se han ejecutado los experimentos. Para facilitar esta tárea en cada experimento se muestran dos gráficas.

La primera de ellas es un cocleograma calculado a partir de una señal acústica de referencia, tomada del "AuditoryToolb". Esta señal de referencia es la mostrada en la figura 5.1.

La segunda es un gráfico en el que apreciamos la variación de la señal de referecia en función del tiempo, para cada una de las bandas obtenidas mediante el filtrado que suponen los distintos modelos auditivos.

5.2.1. Modelo de Lyon

El modelo de Lyon es, de los 3, el que más variaciones permite debido a su elevado número de parámetros. Por ello se han llevado a cabo 3 experimentos con distintos parámetros. Se aporta una explicación de los distintos parámetros en el apéndice B.

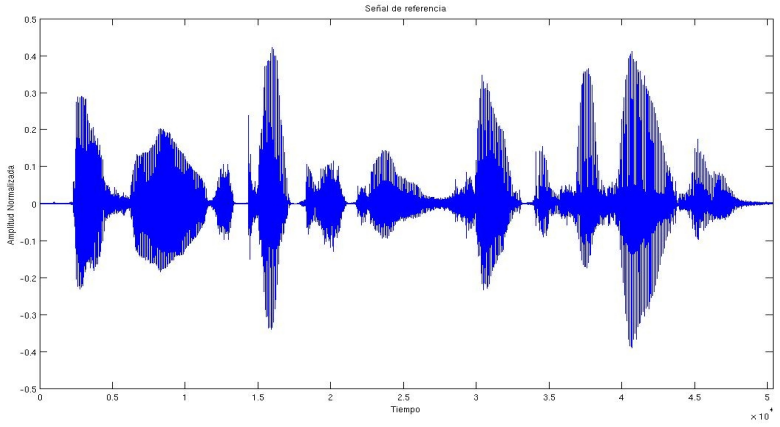

Figura 5.1: Señal de Referencia

Experimento 1

Los distintos parámetros y sus valores son los siguientes:

- **earQ**= 4
- **stepfactor**= 0,25
- **decimation**= 160
- **diff**= 1
- **agcf**= 1
- **tau**= 3

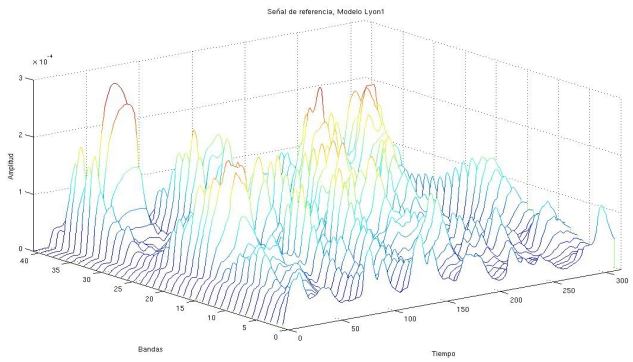

Figura 5.2: Bandas de la Señal de Referencia con el Modelo Lyon1

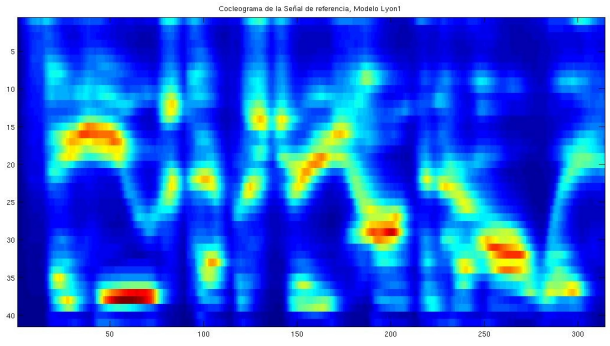

Figura 5.3: Cocleograma de la Señal de Referencia con el Modelo Lyon1

En este experimento se realiza un control automático de ganancia de la señal. Tenemos un factor de calidad de 4, y se realiza un diezmado equivalente a trabajar con ventanas de tamaño 25 ms. La selección concreta de los parámetros da como resultado una análisis con 41 bandas, esto puede apreciarse muy bien en la figura 5.2.

El cocleograma obtenido a partir de este modelo auditivo puede apreciarse en la figura 5.3.

Experimento 2

Los parámetros del modelo han tomado en esta ocasión los siguientes valores:

- **earQ**= 4
- **stepfactor**= 0,25
- **decimation**= 160
- **diff**= 1
- **agcf**= 0
- **tau**= 3

En este caso no se realiza un control automático de ganancia de la señal. Como en el caso anterior tenemos un factor de calidad de 4, y se realiza un diezmado equivalente a trabajar con ventanas de tamaño 25 ms. Las 41 bandas de la salida del modelo pueden apreciarse en la figura 5.4.

El cocleograma obtenido a partir de este modelo auditivo puede apreciarse en la figura 5.5.

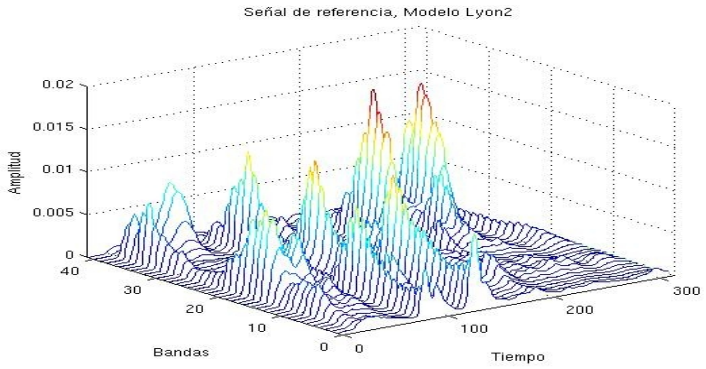

Figura 5.4: Bandas de la Señal de Referencia con el Modelo Lyon2

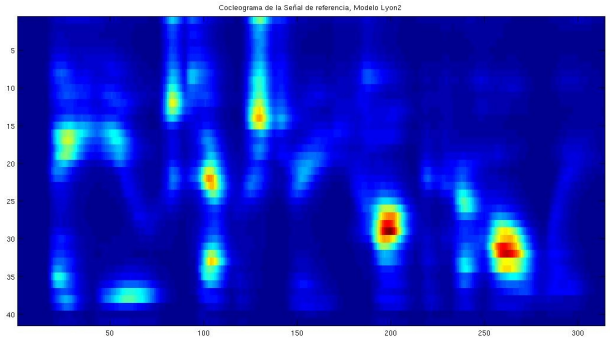

Figura 5.5: Cocleograma de la Señal de Referencia con el Modelo Lyon2

Experimento 3

Para el último experimento con el modelo de Lyon se han escogido los parámetros:

- **earQ**= 8
- **stepfactor**= 0,25
- **decimation**= 160
- **diff**= 1
- **agcf**= 0
- **tau**= 3

Para este experimento tampoco se ha activado el control automático de ganancia de la señal. Tenemos un factor de calidad de 8, y también se realiza

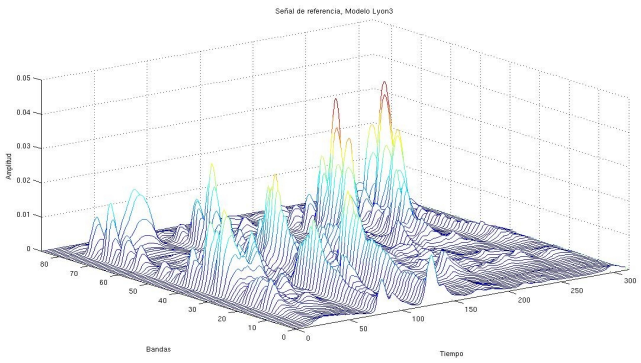

Figura 5.6: Bandas de la Señal de Referencia con el Modelo Lyon3

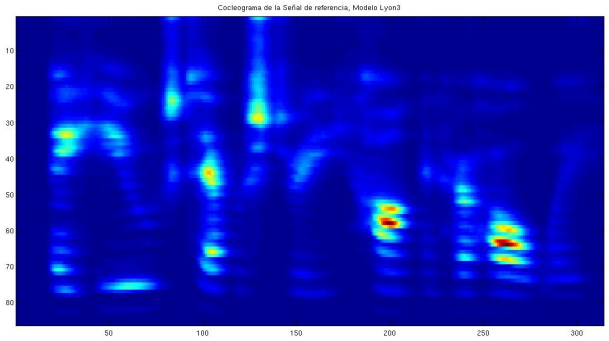

Figura 5.7: Cocleograma de la Señal de Referencia con el Modelo Lyon3

un diezmado equivalente a trabajar con ventanas de tamaño 25 ms. La selección de estos valores permite obtener un análisis con 86 bandas, ver figura 5.6.

El cocleograma obtenido a partir de este modelo auditivo puede apreciarse en la figura 5.7.

5.2.2. Modelo ERB

En el caso del Modelo ERB también se han realizado dos experimentos distintos, uno con 25 bandas y el otro con 45. En ambos casos la frecuencia mínima usada como parámetro ha sido 45Hz.

Existen algunas variaciones en la manera de proceder respecto a los casos anteriores. La implementación del modelo auditivo ERB del "AuditoryToolbox" necesita definir previa y explícitamente el numero de bandas que se desea obtener. Además, no realiza diezmado como Lyon. Por tanto una vez la señal de voz ha sido procesada por el modelo, debe ser filtrada y diezmada. De esta manera podemos seguir trabajando como si estuviéramos con venta-

nas de 25 ms. Además este filtrado elimina información poco relevante para el Reconocimiento de Habla.

Los cocleogramas y las bandas de la señal de referencia que mostraremos en los dos experimentos siguientes ha sido calculado despueés del filtrado y el diezmado.

Experimento 1

Para este experimento hemos calculado un **numero de bandas**= 25.

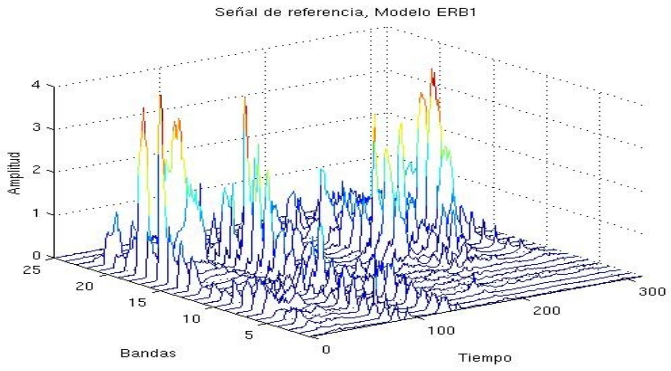

Figura 5.8: Bandas de la Señal de Referencia con el Modelo ERB1

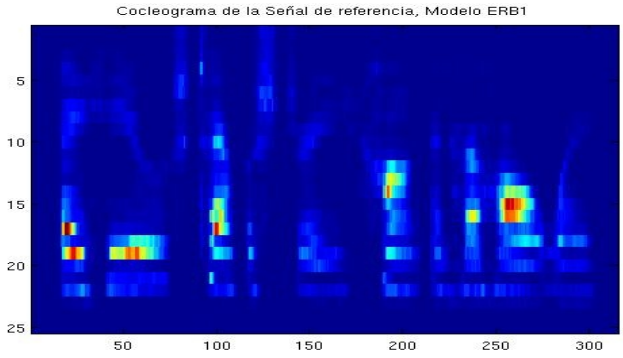

Figura 5.9: Cocleograma de la Señal de Referencia con el Modelo ERB1

Las bandas del modelo y el cocleograma pueden compararse en las figuras 5.8 y 5.9 respectivamente.

Experimento 2

Para este experimento hemos calculado un **numero de bandas**= 45.

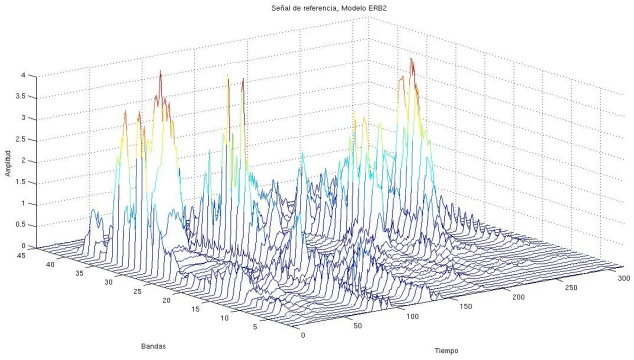

Figura 5.10: Bandas de la Señal de Referencia con el Modelo ERB2

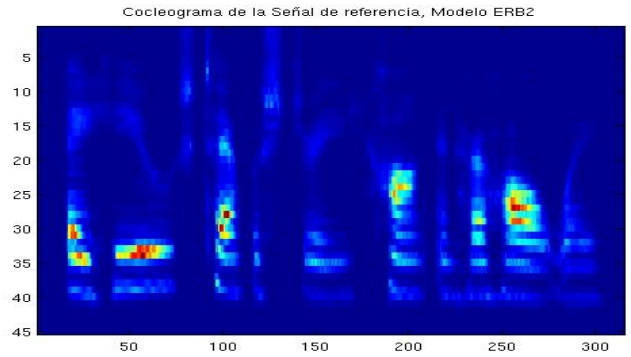

Figura 5.11: Cocleograma de la Señal de Referencia con el Modelo ERB2

En las figuras 5.10 y 5.11 podemos observar las bandas y el cocleograma que este modelo genera para la señal de referencia.

5.2.3. Modelo de Seneff

El último modelo que se ha considerado para los experimentos ha sido el modelo de Seneff. Este modelo carece de parámetros con los que poder jugar en la implementación de Malcolm Slaney. Por tanto sólo tenemos un experimento con este modelo.

El modelo de Seneff siempre trabaja con 40 bandas.

El modelo de Seneff no realizado diezmado, por tanto al igual que ocurría con el modelo ERB es necesario filtrar y diezmar explicitamente.

En las figuras 5.12 y 5.13 se aprecian las bandas y el cocleograma de la señal de referencia una vez procesada, filtrada y diezmada.

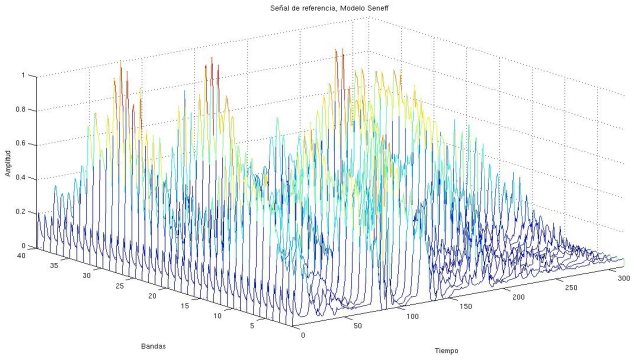

Figura 5.12: Bandas de la Señal de Referencia con el Modelo Seneff

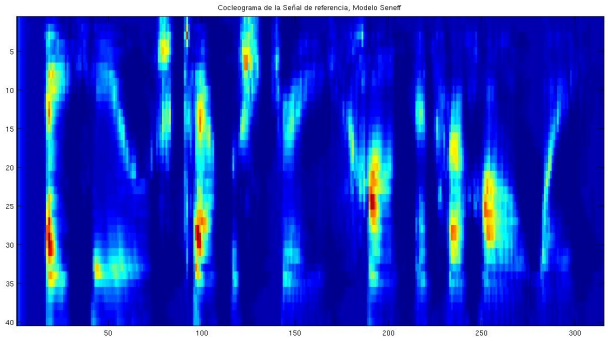

Figura 5.13: Cocleograma de la Señal de Referencia con el Modelo Seneff

5.2.4. Resultados

Como podemos observar a la vista de las figuras de las bandas de las secciones anteriores los modelos utilizados en los distintos experimentos son muy diferentes. Sin embargo, en los diferentes cocleogramas se aprecian fácilmente los mismos patrones, zonas no azul oscuro. Es en estas zonas dónde reside la información fonética, frecuencias de resonancia y formantes. Por tanto aquellos modelos que muestren con más claridad estas zonas deberían de proveer una mejor parametrización y mejores resultados.

La Base de Datos se parametrizó para cada uno de los diferentes modelos y se entrenó el Sistema de Reconocimiento. Se experimentó con contextos de tamaño 5 y 7. Los resultados de test que se obtuvieron se resumen en la tabla 5.3.

También podemos analizar la tabla 5.4, en ella se han inlcuido los porcentajes de mejora (datos positivos) y empeoramiento (datos negativos) que los nuevos modelos obtienen frente a PLP. No se ha incluido en esta ocasión los resultados del test de significancia estadística pues los resultados para todos

	Clean - Clean	Clean - Noisy	Noisy - Noisy
Lyon1 c5	14,5	44,6	41,2
Lyon1 c7	12,4	43,6	39,7
Lyon2 c5	10,5	40,5	24,9
Lyon2 c7	9,9	38,8	23,2
Lyon3 c5	11,2	41,4	24,1
Lyon3 c7	9,6	39	23
ERB1 c5	7,2	54,2	25,8
ERB1 c7	7	52,8	24,7
ERB2 c5	6,9	53,3	24,9
ERB2 c7	6,6	53	23,3
Seneff c5	5,9	39,5	20,9
Seneff c7	5,5	38,3	20,1

Tabla 5.3: Tasa de Error, por palabra. Nuevos Modelos.

% Mejora	Clean - Clean	Clean - Noisy	Noisy - Noisy
Lyon1 c5	-302,77	+16,94	-136,78
Lyon1 c7	-244,44	+18,8	-128,16
Lyon2 c5	-191,66	+24,58	-43,1
Lyon2 c7	-175	+27,74	-33,33
Lyon3 c5	-211,11	+22,9	-38,5
Lyon3 c7	-166,66	+27,37	-32,18
ERB1 c5	-100	-0,93	-48,27
ERB1 c7	-49	+2	-42
ERB2 c5	-91,66	+0,74	-43
ERB2 c7	-83	+1	-34
Seneff c5	-63,88	+26,44	-20,11
Seneff c7	-52,77	+28,67	-15,51

Tabla 5.4: Porcentaje de mejora, respecto a PLP. Nuevos Modelos.

los casos fue el mismo: la probabilidad de la hipótesis nula es **despreciable**.

5.3. Transformada de Cramer

En el tercer gran grupo de experimentos de este Proyecto se ha puesto en práctica lo que comentamos en la sección 3.5.1 de forma teórica.

Para calcular los coeficientes del "cesptrum" se realiza la Transformada de Fourier Inversa del logaritmo del modulo de la Transformada de Fourier. En los experimentos de esta sección se ha calculado la Transformada de Legendre-Fenchel del logaritmo del modulo de la Transformada de Fourier. Lo cual no termina de ser la Transformada de Cramer pero guarda una fuerte semejanza con ella.

Esta nueva forma de parametrizar la señal de voz ha sido evaluada con diferentes modelos auditivos. Al igual que en la sección anterior estos modelos han sido:

- Modelo de Lyon
- Modelo ERB
- Modelo de Seneff

Los parámetros concretos de algunos de los modelos han variado respecto del caso anterior.

Además del modelado (filtrado) auditivo y de la Transformada de Legendre-Fenchel se han realizado otra serie de procesados a la señal. Debido a las implementaciones de los modelos estos procesados han variado ligeramente de unos experimentos a otros. Por este motivo detallaremos en cada apartado el procesamiento realizado a la señal acústica para cada experimento.

5.3.1. Modelo de Lyon

El primer paso de la parametrización es realizar un filtrado paso alto a la señal.

A continuación la señal es modelada. En este caso sólo se ha utilizado una combinación de parámetros, es la siguiente:

- **earQ**= 4
- **stepfactor**= 0,25
- **decimation**= 1

- **diff**= 1
- **agcf**= 1
- **tau**= 3

Esta combinación da lugar a un análisis en 41 bandas.

El siguiente paso consiste en enventanar la señal. De esta manera podremos trabajar con ventanas de 32 ms. Para cada una de estas ventanas se calcula la energía, que será el primero de los elementos del vector de parámetros.

Calculamos la Transformada de Fourier de la señal enventanada. Y el logaritmo del módulo de ésta.

Como comentamos anteriormente, se calcula ahora la Transformada de Legendre-Fenchel. Este cálculo se realiza para cada banda(de las 41 totales). De esta manera, para cada ventana de tiempo tenemos 42 elementos característicos: 1 de energía total y 41 por cada banda analizada. Esto hace un vector de parámetros de longitud 42.

El último paso consiste en realizar un post-enmascaramiento ("post-masking") temporal de los elementos obtenidos en el paso anterior. Para realizar este enmascaramiento se tienen en cuenta tantas ventanas de tiempo como indique el contexto del experimento. En concreto se ha experimentado con contexto 5 y 7.

5.3.2. Modelo ERB

Al igual que en el caso anterior comenzamos filtrando paso-alto la señal de habla. Y enventanamos en tramos de 25 ms. También aquí el primer parámetro característico es la energía total de la ventana, que se calcula en este paso.

Calculamos la Transformada de Fourier de la ventana y el logaritmo del modulo de ésta. Es en este punto dónde filtramos según el modelo ERB elegido. Se han utilizado tres variantes, todas ellas con una frecuencia mínima de 45 Hz:

- **15 bandas**
- **25 bandas**
- **45 bandas**

Para este experimento, debido a la linealidad del modelo ERB, se ha optado por calcular aparte la respuesta al impulso del modelo. Para después

filtrar cada ventana de tiempo con esta respuesta impulsional. Este proceso ahorra tiempo y complejidad a la parametrización.

A continuación calculamos la Transformada de Legendre-Fenchel para cada una de las bandas. Y para terminar se realiza el enmascaramiento. Como en el modelo de Lyon, con contexto 5 y 7.

Los vectores de características tienen un elemento más que el número de bandas de cada modelo. Esto es, 16, 26 y 46 respectivamente.

5.3.3. Modelo de Seneff

Esta parametrización también comienza con un filtrado paso alto para eliminar las frecuencias más bajas.

Se enventana la señal. En este caso con ventanas de 32 ms.

Al igual que en los casos anteriores el primer elemento calculado es la energía total de la ventana.

A continuación la señal enventanada es filtrada por el modelo auditivo. El modelo de Seneff siempre genera 40 bandas, por tanto tenemos 41 elementos característicos. A cada una de estas bandas se les realiza la Transformada de Fourier. Y se calcula el logaritmo del modulo. Con este valor calculamos la Transformada de Legendre-Fenchel.

Como en los otros experimentos el último paso consiste en enmascarar los elementos característicos obtenidos. Con el modelo de Seneff se experimentó con contextos de tamaño 5, 7 y 9.

5.3.4. Resultados

En los apartados anteriores se han realizado una serie de suposiciones sobre la calidad de los distintos modelos bansándose en los cocleogramas mostrados. A la vista de los resultados obtenidos, que se resumen en la tabla 5.5, no podemos confirmar todas estas suposiciones.

Como se hizo en la sección 5.2 mostramos en la tabla 5.6 los porcentajes de mejora y empeoramiento para estos experimentos. Se realizó el test de significación estadística para estos datos. Excepto para el caso 'Seneff c5 clean-noisy' la probabilidad de obtener estos datos fortuitamente es despreciable. En este caso concreto (Modelo de Seneff ajustado con contexto 5) la probabilidad de obtener una mejora del 0,93 % debido a la casualidad es del 35,46 %, bastante alta. Esta probabilidad será tenida en cuenta a la hora de sacar conclusiones.

Las conclusiones de todos los experimentos se encuentran en el siguiente capítulo(6).

	Clean - Clean	Clean - Noisy	Noisy - Noisy
Lyon c5	7,2	60,3	31,1
Lyon c7	8,1	57	27,3
ERB15 c5	5,6	61,8	29,9
ERB15 c7	7,7	61,7	28,2
ERB25 c5	4,8	59,4	26,9
ERB25 c7	6	57,9	25,6
ERB45 c5	4,9	57,4	25,1
ERB45 c7	5,4	57,4	23
Seneff c5	6,8	53,2	25,5
Seneff c7	6,8	51	25,4
Seneff c9	8,8	50	26

Tabla 5.5: Tasa de Error, por palabra. Parametrizaciones con Cramer.

% Mejora	Clean - Clean	Clean - Noisy	Noisy - Noisy
Lyon c5	-100	-12,1	-79,88
Lyon c7	-125	-6,14	-56,89
ERB15 c5	-55.55	-15,08	-71,83
ERB15 c7	-113,88	-14,89	-62,06
ERB25 c5	-33,33	-10,61	-54,59
ERB25 c7	-66,66	-7,82	-47,12
ERB45 c5	-36,11	-6,89	-44,25
ERB45 c7	-50	-6,89	-32,18
Seneff c5	-88,88	+0,93	-46,55
Seneff c7	-88,88	+5,02	-45,97
Seneff c9	-59,09	+6,89	-49,42

Tabla 5.6: Porcentaje de mejora, respecto a PLP. Parametrizaciones con Cramer.

Capítulo 6

Análisis de los Resultados

A partir de los datos obtenidos en los experimentos, podemos llegar a una serie de conclusiones sobre las distintas parametrizaciones utilizadas. Debemos recordar que, como se ha indicado en los capítulos anteriores, las longitudes de los vectores de parámetros son distintas en función del experimento.

Para facilitar su comprensión, las gráficas de barras que se utilizan en este capítulo guardan todas un mismo patrón de diseño.

Los datos referentes al modelo de Lyon aparecen en la gama de verdes, los referentes al modelo ERB en rojos y los del modelo de Seneff lo hacen en azules. Para diferenciar los distintos experimentos que se han llevado a cabo con un mismo modelo, en cada gráfica, tonalidades diferentes de un mismo color indican distintas variantes de un mismo modelo. Por último, las barras que muestran resultados para contexto 5 son lisas, y las de contexto 7 y 9 son rayadas.

Los nombres de los experimentos que aparecen en las leyendas son los utilizados en el capítulo 5.

La primera, y seguramente más evidente, conclusión a la que se ha llegado es que tan sólo se han producido mejoras respecto a "plp" para el caso desajustado (clean train-noisy test). Siendo estas mejoras variables en función del modelo utilizado y el contexto. Las figuras 6.1 y 6.2 nos muestran estos porcentajes de mejora.

Del grupo de experimentos en el que no se introducía la Transformada de Fenchel tan sólo ERB1 c5 no produce mejorías. El resto obtiene muy buenos resutados destacando entre todos el modelo de Seneff.

Por el contrario, aquellos experimentos en los que se ha trabajado con Fenchel no producen mejoría respecto a "plp", a excepción de Seneff. De nuevo Seneff aparece como el modelo con mejores prestaciones. Se recuerda en este punto que según el test de significación estadística el dato de la mejoría de

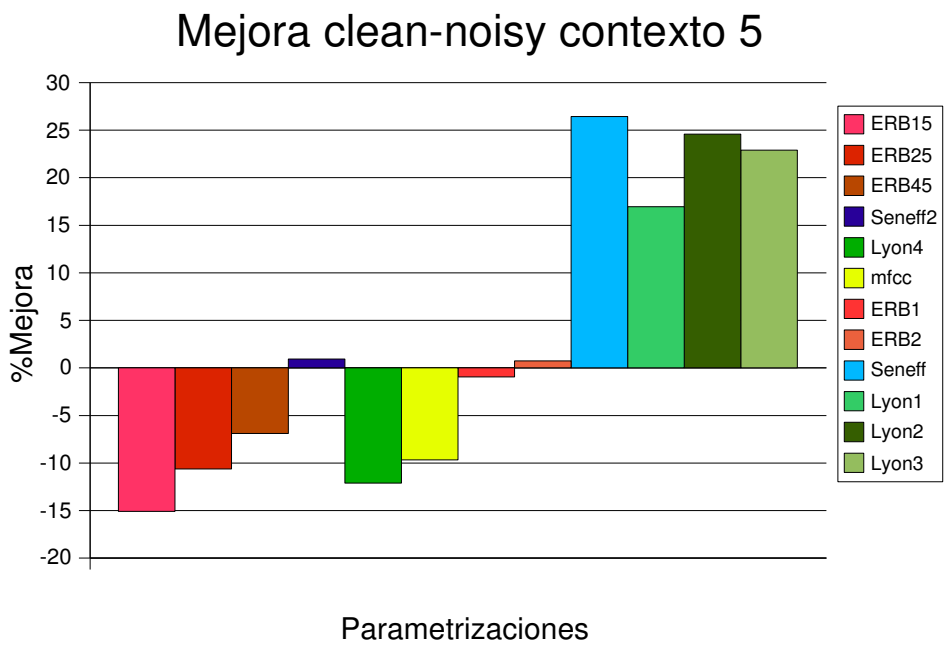

Figura 6.1: Mejora de los experimentos desajustados, contexto 5

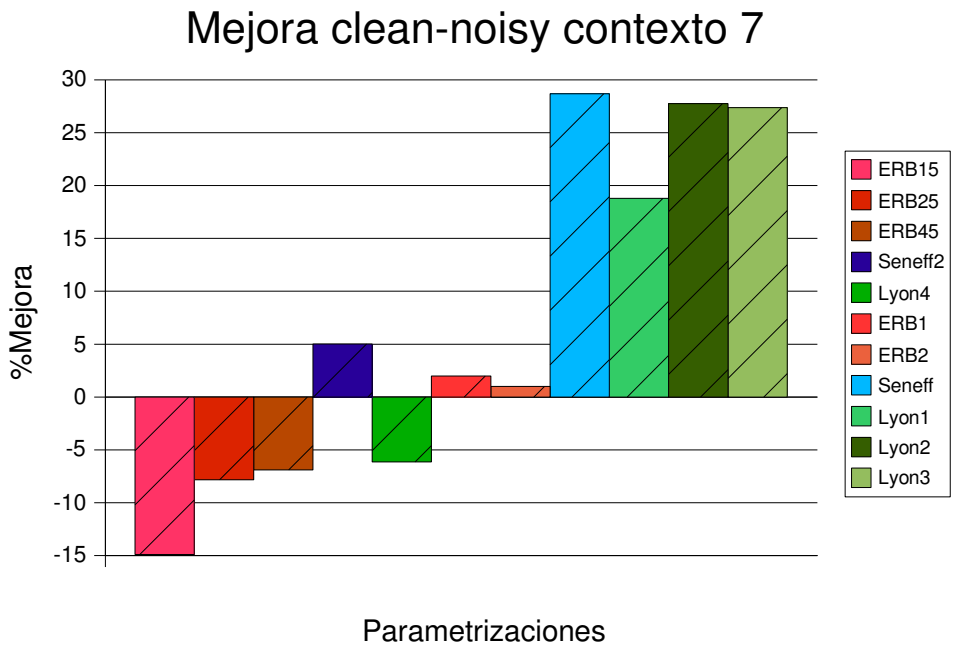

Figura 6.2: Mejora de los experimentos desajustados, contexto 7

Seneff2 c5 sólo es fiable en un 64,54 %.

Es necesario destacar que tanto **Seneff como Lyon2 y Lyon3 mejoran**

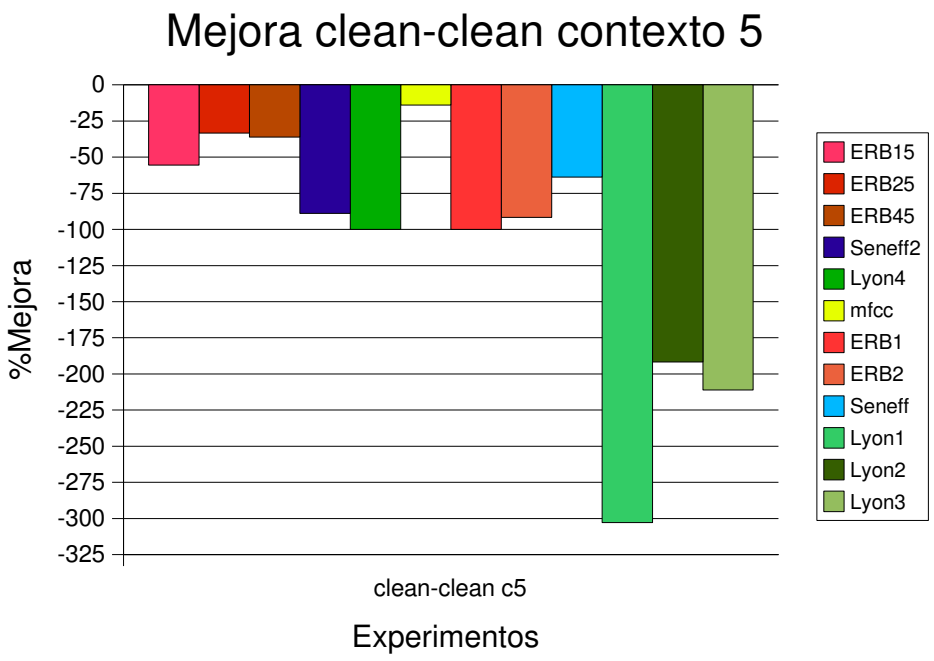

Figura 6.3: Mejora de los experimentos clean-clean, contexto 5

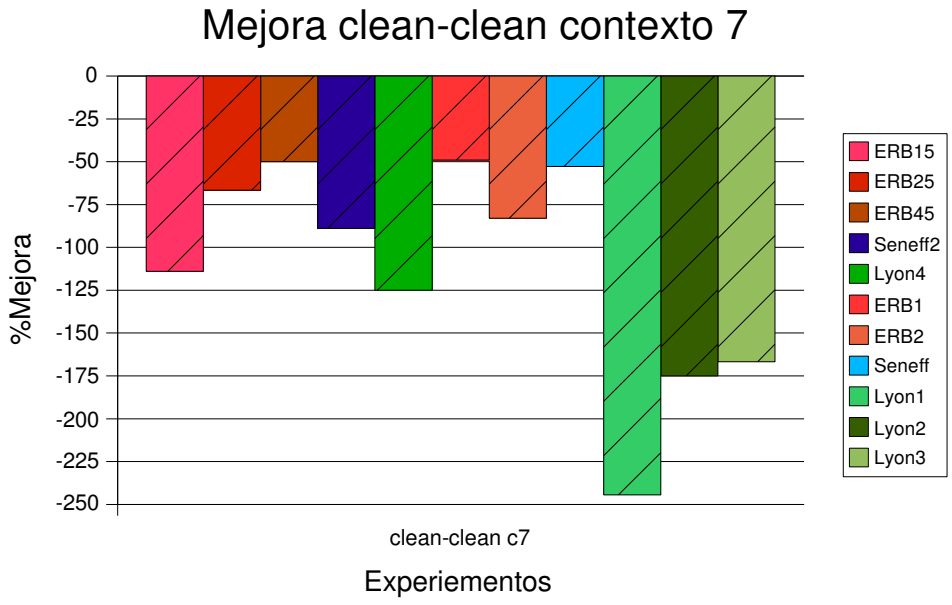

Figura 6.4: Mejora de los experimentos clean-clean, contexto 7

a "plp" en más de un 20 % en el caso desajustado.

Los datos del resto de casos, clean-clean y noisy-noisy, se muestran en

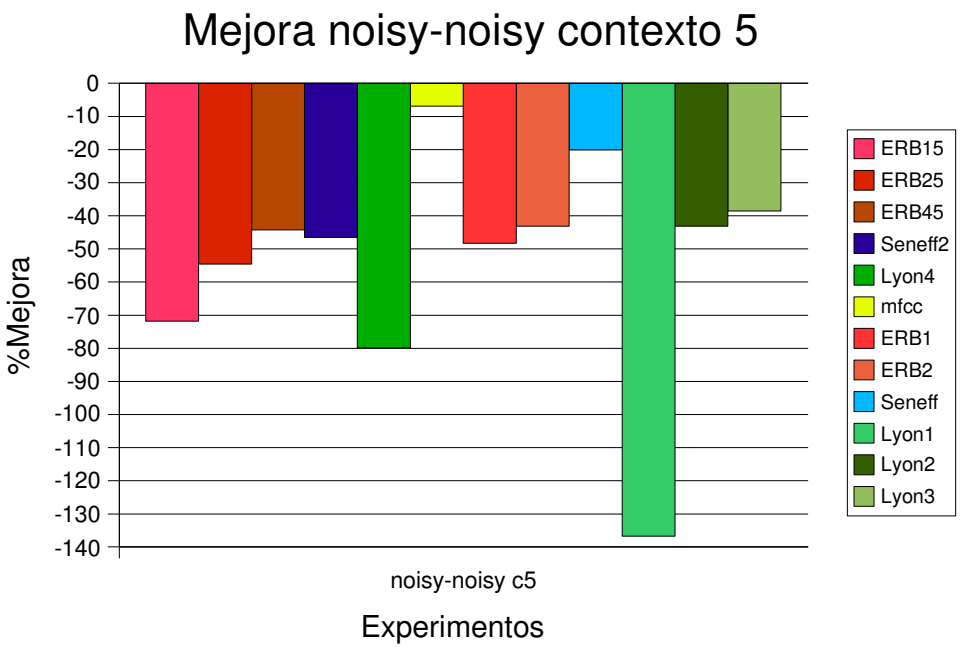

Figura 6.5: Mejora de los experimentos noisy-noisy, contexto 5

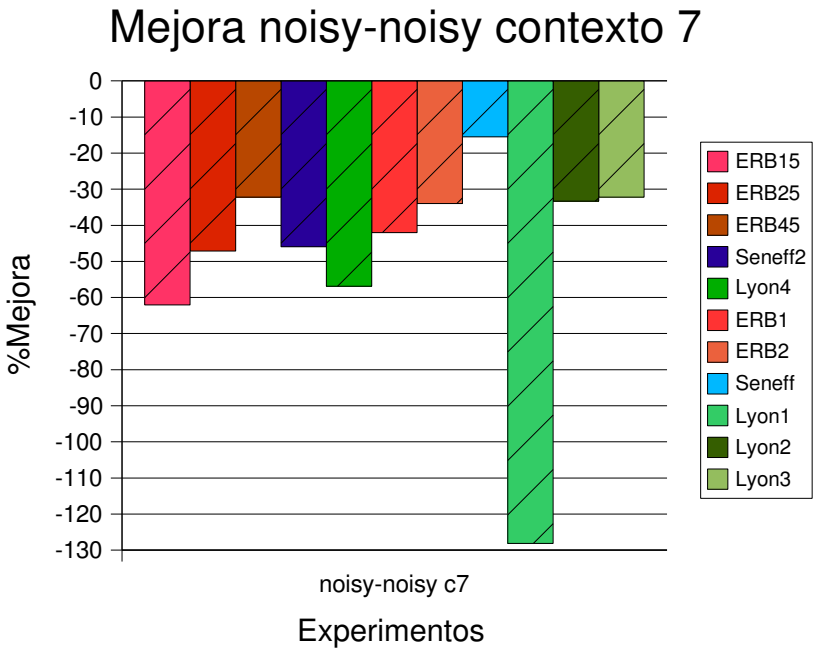

Figura 6.6: Mejora de los experimentos noisy-noisy, contexto 7

las gráficas 6.3, 6.4, 6.5 y 6.6. En estas gráficas se aprecia que los modelos de Lyon dan peores resultados que el resto. Y se aprecia un cierto "empate

técnico" entre ERB y Seneff. Pero todos ellos dan unos porcentajes de mejora mucho peores que los que analizamos en clean-clean.

Otro aspecto importante a tener en cuenta en el Proyecto es la importancia del contexto en la tasa de error. Ilustrando esta cuestión se muestran las figuras 6.7 y 6.8.

Para los experimentos **SIN Transformada de Fenchel** la conclusión es clara: **Aumentar el contexto mejora** las prestaciones. Sin embargo **para el otro caso no** es tan sencillo. Dependiendo del modelo y si el experimento es ajusdado o desajustado los resultados son mejores o peores al aumentar el contexto. No pudiendo apreciar ninguna lógica en este comportamiento.

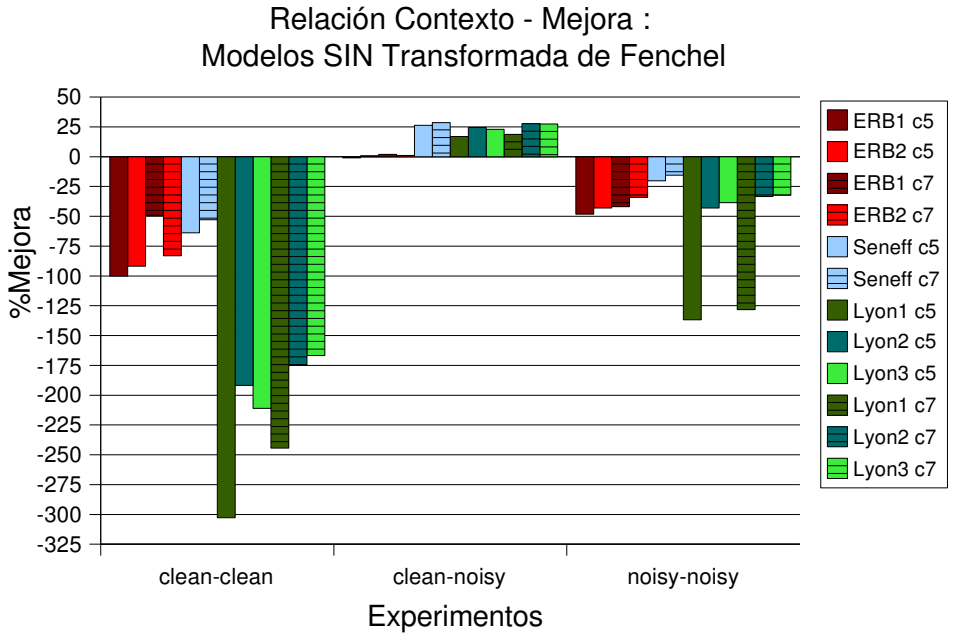

Figura 6.7: Influencia del contexto en la mejora, experimentos sin Transformada de Fenchel

El último aspecto valorado en estas conclusiones ha sido la conveniencia, o no, de hacer uso de la Transformada de Fenchel en los distintos modelos auditivos.

En las figuras 6.9, 6.10 y 6.11 se ha añadido otra codificación. Las barras con borde ancho corresponden a experimentos en los que se incluye la Transformada de Fenchel.

- Modelo de Lyon: podemos ver en la figura 6.9 como el experimento con Transformada de Fenchel esta por detrás del resto en clean-clean y clean-noisy, pero obtiene mejores resultados en noisy-noisy. No obstante el **modelo Lyon1, sin Fenchel, es mejor en todos los casos**.

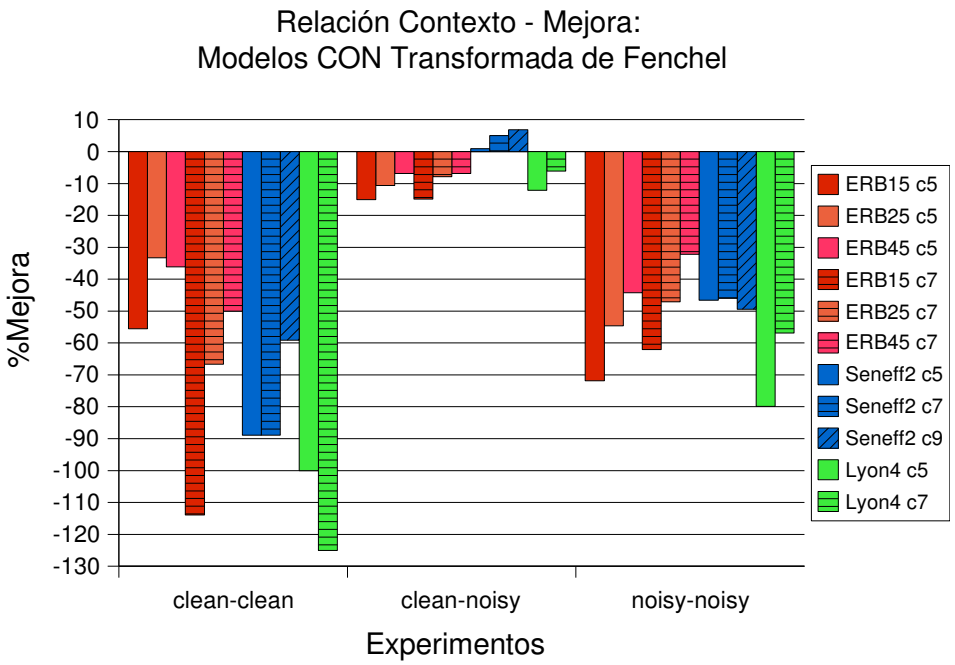

Figura 6.8: Influencia del contexto en la mejora, experimentos con Transformada de Fenchel

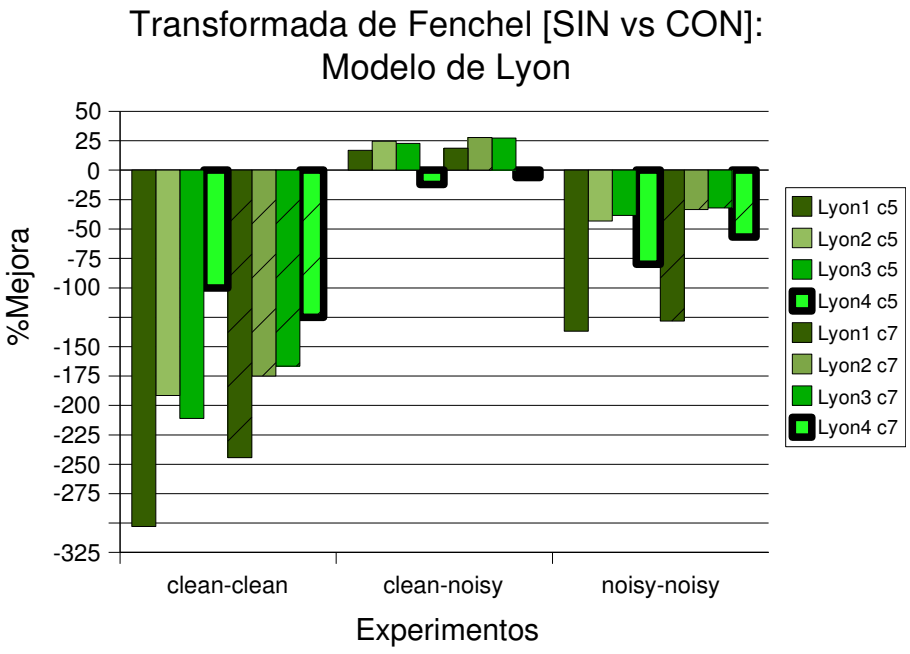

Figura 6.9: Mejora de los experimentos con el modelo Lyon

- Modelo ERB: resulta complicado sacar conclusiones con la figura 6.10.

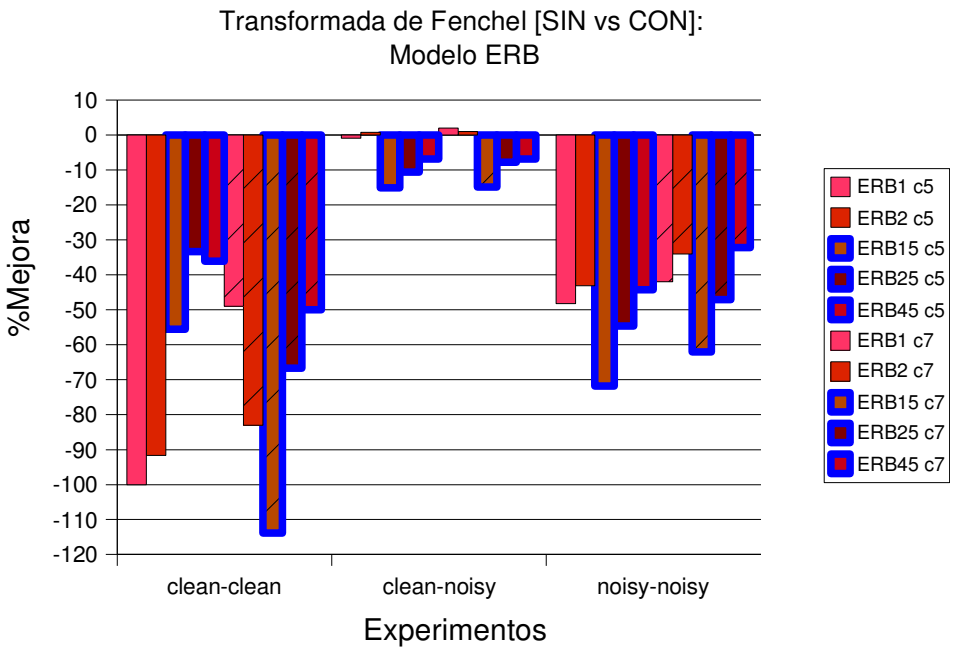

Figura 6.10: Mejora de los experimentos con el modelo ERB

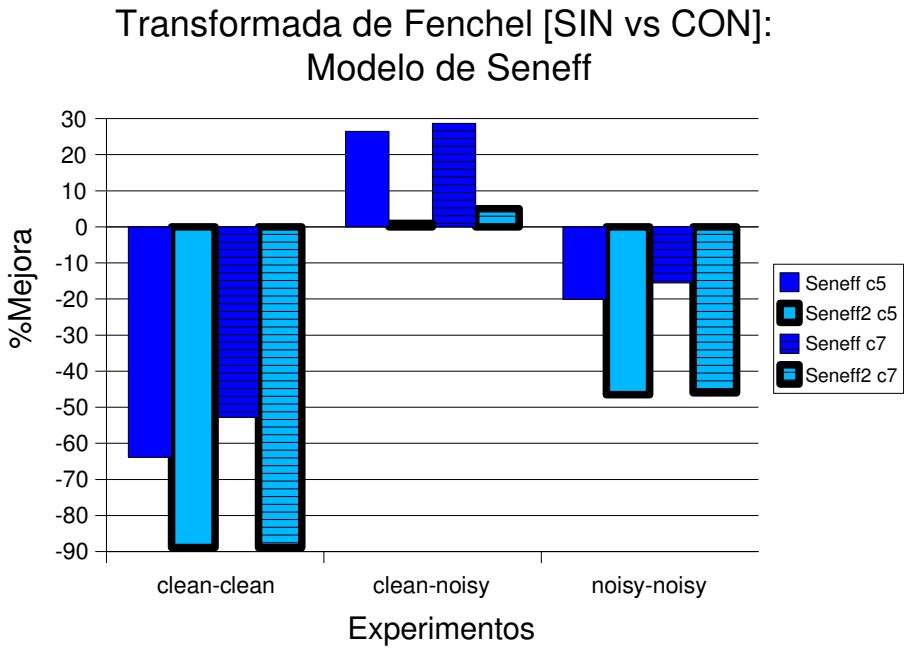

Figura 6.11: Mejora de los experimentos con el modelo Seneff

La Transformada de Fenchel mejora la situación, o no, dependiendo del contexto y el numero de bandas del modelo. Lo que si podemos afirmar

es que a **mayor número de bandas mejores resultados** se obtienen para un mismo tipo de parametrización.

- Modelo de Seneff: la figura 6.11 no deja dudas. Para el **modelo de Seneff es preferible no utilizar la Transformada de Fenchel**.

6.1. Conclusiones Finales

El análisis de los resultados nos permite exponer como conclusiones del Proyecto las siguientes afirmaciones.

La utilización de los nuevos modelos en la parametrización sólo aporta mejoras para el caso clean-noisy.

El modelo de Seneff es el que mejores resultados aporta para entrenamiento limpio y test ruidoso. Llegándose a conseguir casi un 30 % de mejora respecto a plp.

Aumentar el contexto implica un aumento en la calidad para las parametrizaciones sin Transformada de Fenchel. No ocurriendo lo mismo en todos los casos que sí utilizan esta transformada.

Utilizar la Transformada de Fenchel, al menos en la forma en que se ha usado en este Proyecto, empeora la tasa de error del Sistema. Siendo especialmente perjudicial con el modelo de Seneff.

6.2. Líneas Futuras

Este proyecto ha aportado importantes conclusiones sobre la utilización de los diversos modelos auditivos y transformaciones. Gracias a estos datos aparecen nuevas e interesantes cuestiones sobre el tema. Estas cuestiones pudieran ser fruto de nuevos Proyectos e investigaciones. A continuación se apuntan algunas de estas nuevas líneas de trabajo.

- Búsqueda de modelos que permitan parametrizaciones robustas en los casos ajustados.
- Implementaciones de la Transformada de Fenchel que supongan una mejora frente a su no utilización.
- Parametrizaciones basadas en diferentes transformadas que refuercen las prestaciones del modelo de Seneff.

Apéndice A

Instalación del Sistema de Pruebas

El Sistema Automático de Reconocimiento de Habla del que se sirve este Proyecto como Sistema de Pruebas es el "ICSI testbed" desarrollado por el ICSI.

Este sistema de pruebas, que puede descargarse desde la página web del ICSI comprende una serie de herramientas:

- **SPRACHcore**: en concreto se ha utilizado la versión no-gui-2004-08-26.
- **Quicknet**: utilizada la versión v3_20.
- **Dpwelib**: versión 2006-04-19.

La instalación de este conjunto de herramientas no es trivial. A continuación se detallan los problemas, y sus soluciones, que ocurrieron durante la preparación de este Proyecto.

A.1. SPRACHcore

El directorio principal del SPRACcore no dio problemas de instalción. Si los hubo con algunas de sus herramientas:

- **noway**: Es necesario descargar e instalar el patch descargable en *http://www.icsi.berkeley.edu/ dpwe/projects/sprach/noway_patch.tar'*.
- **quicknet**: Numeroso problemas de instalación, se decidió a probar con la versión quicknet3. El resultado está más abajo.

- **feacat**: Requiere dpwelib.

  El configure hace referencia a la versión antigua de quicknet. Hay que sustituir *configname=QNConfig.sh* por *configname=QN3Config.sh* y *configname=qnConfig.sh* por *configname=qn3Config.sh*.

  Los .o no se borran, si se ha compilado con referencias a la versión antigua de quicknet es necesario hacer *make clean* antes de volver a compilar.

- **feacalc**: Igual que para feacat.

- **pfile_utils**: Hay que cambiar en el configure *configname=QNConfig.sh* por *configname=QN3Config.sh*.

  Además en el makefile hay que añadir *LIBS -lintvec*.

A.2. Quicknet3

El configure no da problemas. Sin embargo make:

QN_fir.h:116: error: extra qualification 'QN_InFtrStream_FIR::' on member 'FillDeltaFilt'

QN_fir.h:118: error: extra qualification 'QN_InFtrStream_FIR::' on member 'FillDoubleDeltaFilt'

*make: *** [QN_utils.o] Error 1*

Se soluciona borrando *QN_InFtrStream_FIR::* de las lineas que provocan el error.

A.3. Dpwelib

Fue la parte más problemática de instalar. Cada solución de un problema generaba nuevos problemas, por ello se han incluido todos los problemas que fueron surgiendo.

Configure da la siguiente salida por consola:

...

checking for isatty... yes

checking for strndup... yes

checking whether byte ordering is bigendian... no

checking whether char is unsigned... no

checking system version (for dynamic loading)... ./configure: line 4201: syntax error near unexpected token '('

./configure: line 4201: ' case '(ac_space=' '; set | grep ac_space) 2>&1' in'

Si ejecutamos:

aclocal –acdir 'directorio-del-dpwelib'/dpwelib-2006-04-19

autoconf

./configure

Se soluciona el error. Ahora el problema es con make:

./audIOlinux.c:241:1: error: pasting AUOpen: and SNDCTL_DSP_GETBLKSIZE does not give a valid preprocessing token

Borro la linea 89 de audIOlinux.c (*## #code*)

Finalmente se cambió en el fichero audIOlinux.c la linea en la que aparece *extern int errno* por *#include </usr/include/errno.h>*.

Apéndice B

El "AuditoryToolbox"

Para la imlementación de las parametrizaciones, a excepción de "plp" y "mfcc" se ha utilzado la herramienta matemática MATLAB.

Además de ciertas funciones estándar de sus librerias se ha utilizado una "toolbox" externa específica. El "AuditoryToolbox" desarrollado por Malcolm Slaney[22].

Este conjunto de funciones provee los modelos auditivos utilizados en el Proyecto. Por este motivo, en este anexo, se incluyen las páginas del manual (traducido) de las principales funciones utilizadas.

B.1. LyonPassiveEar

Propósito Calcular respuestas del nervio auditivo usando el modelo coclear pasivo de Lyon.

Sinopsis y=LyonPassive Ear(x, sr, df, earQ, stepfactor, diff, agcf, taufactor)

Descripción Esta función .m calcula la probabilidad de disparo a lo largo del nervio auditivo debida a un sonido de entrada, x, con una tasa de muestreo, sr. Los restantes argumentos son parámetros opcionales de la implementación del modelo coclear y son descritos a continuación. El valor por defecto de cada parámetro se muestra entre paréntesis.

- df(1) Decimation Factor: Cuanto se diezma la salida del modelo.Normalmente el modelo coclear produce una salida por canal en cada instante de tiempo. Estos parámetros permiten a la salida ser diezmada en tiempo (usando un filtro para reducir "aliasing". (Ver también taufactor).
- earQ(8) Quality Factor: El factor de calidad de un filtro es una medida de su ancho de banda. En este caso mide el cociente del ancho

de cada filtro paso banda a 3dB bajo el máximo. Normalmente, filtros de bandas críticas tienen una Q alrededor de 8. Menores valores de earQ significan filtros cocleares más anchos.

- stepfactor(0.25) Filter stepping factor: Cada filtro de un banco de filtros esta solapado por una fracción fijada mediante este parámetro. El valor por defecto esta dado por earQ/32.
- differ(1) Channel Difference Flag- Los canales de filtros adyacentes pueden ser restados para mejorar la respuesta en frecuencia del modelo. Este parámetro es un flag; valores distintos de cero indican que las diferencias entre canales deben ser calculadas.
- agcf(1) Automatic Gain Control Flag- Un control automático de ganancia se utiliza para modelar la adaptación neuronal. Este flag cambia el mecanismo de adapatación [on-off].
- taufactor(3) Filter Decimation Tau Factor: Cuando la salida del modelo coclear esta diezmada, un filtro paso bajo es aplicado en cada canal para reducir el contenido en altas frecuencias y minimizar el "aliasing". La constante temporal del filtro (tau) viene dada por el decimation factor multiplicado por este argumento. Valores mayores de taufactor implican que menos información de altas frecuencias pase.

Observe que la función resetea el estado del filtro cada vez que es ejecutada. El estado inicial del filtro AGC es cero, por lo que el modelo coclear es muy sensitivo a los sonidos iniciales.

B.2. MakeERBFilters

Propósito Diseñar los filtros necesarios para implemenar un modelo coclear ERB.

Sinopsis fcoefs=MakeERBFilters(fs,numChannels,lowFreq)

Descripción Esta función computa los coeficientes de los filtros para un banco de filtros "Gammatone". Estos filtros fueron definidos por Patterson y Holdworth para simular la cóclea. El resultado es devuelto como un array de coeficientes de filtros. Cada columna del array de filtros contiene los coeficientes para cuatro filtros de segundo orden. La función de transferencia para estos cuatro filtros comparte el mismo denominador (polos) pero tiene diferentes numeradores (ceros). Todos estos coeficientes estan ensamblados en un vector que la función ERBFilterBank puede tomar aparte para implementar el filtro.

El banco de filtros contiene numChannels canales que se extienden desde la mitad de la frecuencia de muestreo, fs, hasta lowFreq.

B.3. SeneffEar

Propósito Implementa fases I y II del Modelo Auditivo de Seneff.

Sinopsis y=SeneffEar(x, fs [, plotChannel])

Descripción Esta función implementa la Fase I (Banco de Filtros de Bandas Críticas) y la Fase II (Modelo Synapse Hair Cell) del modelo Auditivo de Seneff. Esta rutina convierte una señal de entrada, x, en un array de "formas de onda detalladas de la respuesta probabilística a ciclos individuales del estímulo de entrada". Este modelo esta "basado en propiedades del sistema auditivo humano. Un banco de filtros de banda crítica define el análisis espectral inicial. Las salidas de los filtros son procesadas por un modelo de las fases no-lineales de transducción en la cóclea, las cuales se encargan de características como la saturación, la adaptación y el 'forward masking'. Los parámetros del modelo fueron ajustados para coincidir con los resultados experimentales de fisiología del sistema auditivo."

B.4. Sobre el código

Para terminar este apéndice se exponen dos problemas que surgieron al hacer uso de estas funciones:

- Para algunas de las funciones de la libreria es necesario tener declarada la variable LD_PRELOAD con el valor /lib/libgcc_s.so.1
- Existe un fallo en la programación de la función SeneffEarSetup. Es necesario cambiar la declaración de los contadores de los dos últimos bucles.

Apéndice C

Presupuesto del Proyecto

En este anexo se presentan justificados los costes globales de la realización de este Proyecto Fin de Carrera. Tales costes, imputables a gastos de personal y de material, se pueden deducir de los cuadros C.1 y C.2.

En el cuadro C.1 se muestran las fases del proyecto y el tiempo aproximado para cada una de ellas. Así pues, se desprende que el tiempo total dedicado por el proyectando ha sido de 1 050 horas, de las cuales aproximadamente un 10 % han sido compartidas con el tutor del proyecto, por lo que el total asciende a 1 155 horas. Considerando una tarifa de 63,57 €/hora, el coste de personal se sitúa en 73 423,35 €.

En el cuadro C.2 se recogen los costes de material desglosados en equipo informático, local de trabajo, documentación y gastos varios no atribuibles (material fungible, llamadas telefónicas, desplazamientos...). Ascienden, pues, a un total de 3 720 €.

A partir de estos datos, el presupuesto total es el mostrado en el cuadro

Fase	**Descripción**	**Núm. de horas**
1	Documentación	420 horas
2	Análisis e Instalación de la base de datos	20 horas
3	Análisis e Instalación del Sistema de Pruebas	240 horas
4	Codificación	80 horas
5	Experimentación y Toma de Resultados	150 horas
6	Redacción de la memoria	140 horas

Tabla C.1: Fases del Proyecto
Listado de las diferentes fases del proyecto y de las horas dedicadas a cada una de ellas

Ordenador portatil de gama media	900 €
Local (durante 12 meses, con un coste de 145 €/mes)	1 740 €
Documentación	180 €
Gastos varios	900 €

Tabla C.2: Costes de material
Costes imputables

Concepto	**Importe**
Costes personal	73 423,35 €
Costes material	3 720 €
TOTAL	77 143,35 €

Tabla C.3: Presupuesto
Presupuesto Total del PFC

C.3.

Apéndice D

Abreviaturas y Siglas

A continuación aparece una recopilación de las abreviaturas y siglas utilizadas en el Proyecto.

AGC: "Automatic Gain Control"; Control Automático de Ganancia

ANN: "Artificial Neuronal Network"; Red Neuronal Artificial

ASR: "Automatic Speech Recognition"; Reconocimiento Automático de Habla

CMS: "Cepstral Mean Subtraction"

ERB: "Equivalent Rectangular Bandwith"; Ancho de Banda Rectangular Equivalente

HMM: "Hidden Markov Model"; Modelo Oculto de Markov

HWR: "Half Wave Rectifier"; Rectificador de Media Onda

ICSI: "International Computer Science Institute"

IHC: "Inner Hair Cells"; Células Ciliadas Internas

LPC: "Linear Predictive Coefficients"

MFCC: "Mel Frecuency Cepstral Coefficients"

MLP: "Multi-Layer Perceptron"; Perceptrón Multicapa

OHC: "Outer Hair Cells"; Células Ciliadas Externas

PLP: "Perceptual Linear Prediction"

RAH: "Reconocimiento Automático del Habla"

RASTA: "RelATive SpecTrAl"

SFSA: "Stocastic Finite State Automaton"; Autómata de Estados Estocásticos Finitos

SNR: "Signal Noisy Ratio"; Relación Señal a Ruido

TC: "Transformada de Cramer"

TF: "Transformada de Fourier"

TLS: "Teoría Lineal de la Señal"

TMS: "Teoría Morfológica de la Señal"

TS: "Transformada Slope", "Transformada de la Pendiente"

TZ: "Transformada Z"

Bibliografía

[1] http://cslu.cse.ogi.edu/corpora/isolet. Technical report.

[2] http://en.wikipedia.org/wiki/Filter_bank. Technical report.

[3] http://es.wikipedia.org/wiki/Conjugada_convexa. Technical report.

[4] http://es.wikipedia.org/wiki/Transformada_de_Legendre. Technical report.

[5] http://faculty.vassar.edu/lowry/binomialX.html. Technical report.

[6] http://www2.pd.istc.cnr.it/pages/asr-seneff.htm. Technical report.

[7] http://www.icsi.berkeley.edu/Speech/papers/eurospeech05-onset/isolet. Technical report.

[8] http://www.mmk.ei.tum.de/demo/imagedb/bitmaps/hmm.gif. Technical report.

[9] Leonardo C. Araujo and company. A Brief History of Auditory Models. Technical report, CEFALA and CPDEE.

[10] Herve Bourlard and Nelson Morgan. Hybrid HMM/ANN Systems for Speech Recognition: Overview and New Research Directions. Technical report, IDIAP, Martigny, Switzerland. Intl. Comp. Science Institute, Berkeley, CA. UC Berkeley, Berkeley, CA, 1998.

[11] Herve A. Bourlard and Nelson Morgan. *Connectionist Speech Recognition.* Kluwer Academic Publishers Group, 1994.

[12] Bernhard Burgeth and Joachim Weickert. An Explanation for the Logarithmic Connection between Linear and Morphological System Theory. Technical report, Mathematical Image Analysis Group, Faculty of Mathematics and Computer Science, Building 27. Saarland University, Germany, 2004.

[13] Pierre Buser and Michel Imbert. *Audition.* Massachusetts Institute of Technology, 1992.

[14] Leo Dorst and Rein van den Boomgaard. Morphological Signal Processing and the Slope Transform. Technical report, Department of Mathematics and Computer Science University of Amsterdam, 1993.

[15] Sharma et al. Feature extraction using non-linear transformation. Technical report, ICASSP, 2000.

[16] Ben Gold and Nelson Morgan. *Speech and Audio Signal Processing.* John Wiley and sons, 2000.

[17] Hynek Hermansky and Nelson Morgan. RASTA Processing of Speech. Technical report, IEEE, 1994.

[18] Ludek Muller Josef Psutka and Josef V. Psutka. Comparison of MFCC and PLP Parameterizations in the Speaker Independent Continuous Speech Recognition Task. Technical report, University of West Bohemia, Czech Republic, 2001.

[19] Daniel Jurafsky and James H. Martin. *Speech and Language Processing.* Prentice Hall, 2000.

[20] L.L. Langley. *Elementos de Fisiologia.* Editorial Acribia, 1973.

[21] Malcolm Slaney. An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank. Technical report, Apple Computer, Inc, 1993.

[22] Malcolm Slaney. Auditory Toolbox version 2. Technical report, Interval Research Corproation, 1998.

[23] Malcolm Slaney. Lyon's Cochlear Model. Technical report, Apple Computer, Inc, 1998.

[24] H. Steeneken and F. Geurtsen. Description of the RSG-10 noise database. Technical report, TNO Institute for Perception, 1988.